Development of a Dynamic Linear Model Procedure for Quantifying Long-term Trends in Atmospheric Time Series

A Thesis Submitted to the

College of Graduate and Postdoctoral Studies

in Partial Fulfillment of the Requirements

for the degree of Master of Science

in the Department of Physics and Engineering Physics

University of Saskatchewan

Saskatoon

By

Curtis Puetz

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Physics and Engineering Physics
163 Physics Building, 116 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5E2
Canada

Or

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

With satellite remote sensing instruments, global data records of various atmospheric species, spanning considerable periods of time, have been produced. These data provide insight into atmospheric processes and the evolution of our atmosphere. Statistical analysis on them is essential. One thing in particular that we often wish to know about is the long-term trend in a species concentration on the order of decades. This is important because it allows us to monitor changes in our atmosphere. Changes that can be traced back to human activity, giving us feedback on how we are affecting the atmosphere, or changes from natural phenomena, such as volcanic eruptions.

In this thesis, a statistical procedure is developed for modelling atmospheric remote sensing data records, with particular emphasis placed on the ability to extract accurate and informative information about the long-term trend. Procedures operating on the same principals have been used in the past for time series analysis in general. For example, on economic time series, as well as on atmospheric remote sensing data records, or just any atmospheric data. In this thesis, we show the theory behind the procedure in detail as well as describe how to implement and use it in practice. This is done with the intent of making the rather complicated procedure more accessible so that it can become more adopted by scientists working with atmospheric remote sensing data if desired, and compared to current methods for obtaining long-term trends.

For an example application of this procedure, we apply it to a stratospheric ozone data record that extends from 1984 to present (2019). Ozone is a species that is of considerable interest since we know without a doubt that the changing chlorine situation in the atmosphere due to human activity has a significant effect on it, and because of its importance in absorbing ultraviolet radiation, which can seriously harm life on the Earth. The results we give paint a detailed picture of the long-term trends in stratospheric ozone concentration in the 65°S to 65°N latitude region.

Acknowledgements

I would like to show my appreciation for the opportunity of performing this work with the Atmospheric Research Group at the University of Saskatchewan. This opportunity has truly enriched my education. I would also like to thank both the Institute of Atmospheric Studies at the University of Saskatchewan and the NSERC Create program for the financial support I have received throughout this degree.

My sincerest thanks go to my supervisor, Dr. Doug Degenstein, for his guiding direction in completing this work and for his always open door. My thanks also go to everyone in the Atmospheric Research Group, especially Chris Roth, Taran Warnock, Dr. Adam Bourassa, and Dr. Daniel Zawada, for taking the time to provide suggestions and feedback for my work, as well as just listening to my ideas.

Lastly, I would like to thank my family for their constant unwavering support.

Contents

List of Figures

List of Abbreviations

DLM	Dynamic Linear Model
MLR	Multiple Linear Regression
SAGE	Stratospheric Aerosol and Gas Experiment
OSIRIS	Optical Spectrograph and InfraRed Imaging System
OMPS	Ozone Mapping and Profiler Suite
SOO	SAGE II/OSIRIS/OMPS
WMO	World Meteorological Organization
LOTUS	Long-term Ozone Trends and Uncertainties in the Stratosphere
MCMC	Markov Chain Monte Carlo
MZM	Monthly Zonal Mean
QBO	Quasi-biennial Oscillation
ENSO	El Niño-southern Oscillation
AOD	Aerosol Optical Depth
OLS	Ordinary Least Squares
GLS	Generalized Least Squares
BLUE	Best Linear Unbiased Estimator
FGLS	Feasible Generalized Least Squares
RLS	Recursive Least Squares
MLE	Maximum Likelihood Estimate
MSE	Mean Squared Error
MMSE	Minimum Mean Squared Error

1 Introduction

The purpose of this research is on the development of a dynamic linear model (DLM) procedure to be used by the Atmospheric Research Group at the University of Saskatchewan, and potentially by the larger atmospheric satellite remote sensing community that this group is a part of. The work has resulted in the development of software that can execute this procedure on desired atmospheric data. The primary purpose of this thesis is to describe the procedure in detail and to give theoretical background on the statistical methods it utilizes.

The application that prompted the development of this procedure is the problem of quantifying temporal trends in stratospheric ozone concentration from satellite remote sensing data records. Since the Antarctic ozone hole scare in the 1980s, which we discuss later in this introduction, the monitoring of ozone concentration trends has become of great concern. Historically, a multiple linear regression (MLR) procedure known as Prais-Winsten estimation has been used to quantify trends with these satellite data records. For which, secondary to the primary purpose of this thesis, this thesis provides detail of the underlying theory and explanation of the algorithm, making this subject more easily accessible to the reader. However, the Prais-Winsten estimation seems to be on its way out for this stratospheric ozone problem, and procedures using DLMs are starting to become adopted with results having already been published (Ball et al., 2017; Ball et al., 2018; Laine et al., 2014). So, in this thesis, we work closely with this example application of stratospheric ozone. But, we importantly note that the developed DLM procedure of this thesis work, which we will commonly refer to in this thesis as just the "DLM procedure", may also be well suited for other atmospheric time series and may even be underused in this area of study. The importance of atmospheric ozone for our planet and the history of its measurement will now be discussed.

Ozone, despite making up a small portion of Earth's atmosphere, has a disproportionately large impact on the atmospheric state and life of earth. Its peak concentration occurs in

the equatorial lower stratosphere and is only about 10 ppmv. For solar radiation in the wavelength range of 200 to 300 nm, ozone is the primary absorber. This absorption is the main driver of the increase in temperature that defines the earth's stratosphere, which is not a characteristic of the Mars or Venus atmospheres. But perhaps more importantly, is the role it plays in protecting life on earth, since solar radiation in this wavelength range can interact with and alter DNA as well as affect normal life processes in many organisms (Farman et al., 2010). For these reasons, the monitoring of ozone in the atmosphere has been of interest, with probably the most sought after statistic being the temporal trend in ozone concentration across the globe. Herein "ozone trend" or "trends in ozone" will refer to long-term changes in ozone concentration with time.

Measurements of ozone in the atmosphere began in the 1920s with the use of ultraviolet spectrometers (Rodgers, 2000). These devices were situated on the ground and originally could only give an indication of the total amount of ozone in a vertical column of the atmosphere. Eventually, ground-based techniques were adapted to allow for information on the vertical distribution of ozone to be extracted (Dobson, 1968). Shortly after, when humanity became capable of sending man-made objects to space, the possibility of putting atmospheric remote sensing instruments on satellites became a reality. The advantage of this over ground-based instruments is that measurements can be made across the entire globe in a relatively short period of time, with the same instrument.

To assess the trends in stratospheric ozone over a long period, merged data records from instruments, typically situated on satellites, that measure vertical distributions of ozone have been used. The data records are said to be merged in that the data that forms them is collected from multiple instruments, rather than one. Given the utilized instruments are operational in different time periods, the result is data that spans a longer period of time than any data record formed from an individual instrument could. This makes the data record much more suitable for the assessment of long-term ozone trends.

The Stratospheric Aerosol and Gas Experiment (SAGE) II was a very successful early NASA satellite instrument that measured ozone. It had an operational lifetime from 1984 to 2005. After about the year 2000 a significant number of ozone retrieving satellite instruments became operational (Petropavlovskikh et al., 2019). Two of such instruments are the Optical

Spectrograph and InfraRed Imaging System (OSIRIS), operational since 2001 to present, and the limb sensor on the Ozone Mapping and Profiler Suite (OMPS), operational since 2011 to present. Using data from these three stratospheric ozone retrieving instruments, a merged data record can be obtained which extends from 1984 to the present. We call this the SAGE II/OSIRIS/OMPS (SOO) data record. This data record is produced at the University of Saskatchewan, and we use it frequently throughout this thesis.

The observed ozone trends to date are very well summarized in the World Meteorological Organization's (WMO) Scientific Assessment of Ozone Depletion: 2018 (WMO, 2018) and the Long-term Ozone Trends and Uncertainties in the Stratosphere (LOTUS) initiatives 2019 report (Petropavlovskikh et al., 2019). A brief overview is given here.

A considerable drop-off of ozone concentration in the Antarctic spring season, starting around 1975, was observed with ground-based instruments and first published famously in 1985 (Farman et al., 1985). Satellite data subsequently confirmed the findings and that the entire Antarctic polar vortex region was affected. This phenomenon is now referred to as the "ozone hole". It was later found that chlorofluorocarbons, which were widely used in refrigerants and solvents at the time, were mainly responsible for this. An agreement called the Montreal Protocol was quickly made by the United Nations in 1987 to ban the use of these substances. This protocol is considered to be very successful, and it is now generally agreed upon that the ozone hole has neither gotten larger nor smaller since about the year 1990. In fact, the WMO reported for the first time in 2018 that a small statistically significant decline of the Antarctic ozone hole size and depth from the year 2000 had been identified.

For other regions across the globe, trends in stratospheric ozone are examined with merged ozone data records like SOO among others (Bourassa et al., 2018; Damadeo et al., 2014; Frith et al., 2014; Gebhardt et al., 2014; Kyrölä et al., 2013; Laeng et al., 2017; Nair et al., 2013; Sofieva et al., 2017; Steinbrecht et al., 2017). The LOTUS initiative has been developed in recent years with the goal of examining the statistical significance of the derived trends. The consensus shows that between about 1984 and 1997 a significant decline in ozone concentration took place in the upper stratosphere (to a larger degree in the middle to high latitudes than the tropics), and a smaller, but still statistically significant, degree of incline is observed in the same regions from the year 2000 to 2016. In the lower stratosphere, trends

are small if not zero and are not statistically significant. Although, negative trends are observed in the equatorial region of the lower stratosphere in many data records. In terms of total ozone columns (integrating a region of atmosphere over altitude), negative trends were observed in some regions of the globe before around 1997 to 2000, but no positive trends afterward are statistically significant. This is the case because upper stratospheric ozone is only a small portion of the total ozone column.

Quantifying trends in ozone is complicated by its variation caused by various ozone influencing phenomena. For example, in some regions of the atmosphere, ozone is dependent on the quasi-biennial oscillation and in others on the sun's solar cycle. These ozone influencing phenomena are regularly called either, predictors, regressors, proxies, or explanatory variables for ozone. The statistical procedures shown in this thesis, using DLMs or MLR models, must account for these. MLR models do this by assessing the strength of the relationship between ozone and these influencing phenomena. In doing so, it can be thought that the model effectively "extracts out" the signals of the influencing phenomenon from the data. With these signals extracted out, theoretically the only signals left in the data are the long-term trend signals. The DLMs we use also extract out the signals of the ozone influencing phenomena in an essentially equivalent way.

To infer trends with the leftover signal, for the MLR model one method is to construct additional predictors that are linear/straight lines over a period of time. The MLR model will then asses the strength of ozone's relationship to these predictors also, and from this, an estimation of the assumed to be linear trend in ozone over a region of time is obtained. The DLMs we use take a different approach. They do not constrict the shape of the trend to be linear over a wide region of time, instead, they estimate the trends uniquely at any instant in time. Often, the construction of the linear predictors in the MLR is not well suited for the ozone data. So, with the MLR model having no better alternative the DLM gains advantage in this respect. The developed DLM procedure of this thesis work also does a better job of estimating the statistical significance of trends, where we show in this thesis that the Prais-Winsten procedure underestimates these more so than the DLM procedure. For these reasons, it is concluded at the end of this thesis that the DLM procedure is more suited for ozone trend analysis than the Prais-Winsten estimation procedure. Again, the

development and implementation of this DLM procedure is the primary result of this thesis work.

In Chapter 2, we show how the SOO data record is created, give theory behind MLR and Prais-Winsten estimation, and give the Prais-Winsten estimation results for the SOO data record. Again, the Prais-Winsten estimation is the primary way in which stratospheric ozone trends are quantified to date. In Chapter 3, an introduction to the topic of Markov Chain Monte Carlo (MCMC) is given. MCMC finds itself as an essential component of the DLM procedure that we develop in this thesis. In Chapter 4, the DLM is introduced. An introductory build up into the topic of DLMs is provided and we go deep into the underlying theory behind the DLM estimation equations. We also provide instruction on how to construct a suitable DLM for any given problem. In our case, we show the example of modelling stratospheric ozone. In Chapter 5, the DLM procedure will finally be presented. This brings all the topics of this thesis together and describes what is implemented in the software that goes along with this thesis work. Lastly, in Chapter 6, we show the results of the application of the DLM procedure to the SOO data record. These results paint a detailed picture of the long-term ozone trends in the stratosphere.

We note that one of the goals of this thesis is to provide as much detail as possible about the theoretical background underpinning the two statistical procedures. The idea is for this thesis to serve as a reference document for this material. There are some longer sections that contain this information, and much of the details are put into short appendices to make this thesis more readable. So, since this document is rather long for a master's thesis, we note here that these sections of the thesis can be skipped without losing any major context for future sections. Specifically, the parts that can be skipped are Section 4.1, Section 4.5, and all of the appendices. Skipping these would bring the required reading down from 234 pages to a more manageable 124 pages. Further to this, if this is the approach, the derivations and theoretical background details in the background chapter can also be glossed over.

2 Background

In this background chapter, we show how the SOO data record is constructed and present the Prais-Winsten estimation procedure. In Section 2.1 we give a brief overview of the three instruments that make up the SOO data record, describe what format the SOO data record is reported in, and describe how data between the instruments are merged. In Section 2.2 we give all the ozone influencing phenomena that are used as predictors in the MLR model. In Section 2.4 we give theory for the MLR model and Prais-Winsten estimation, and in Section 2.5 we give the results of the application of the Prais-Winsten estimation procedure to the SOO data record. These results are the primary way in which stratospheric ozone trends are quantified to date.

2.1 The SAGE II/OSIRIS/OMPS Merged Data Record

SAGE II was a NASA instrument onboard the Earth Radiation Budget Satellite that was operational from 1984 to 2005. It used the limb occultation approach, viewing sunsets and sunrises in wavelengths from approximately 0.2 μm to 1 μm. Its data is presented with a spatial sampling of 0.5 km in altitude.

OSIRIS is a Canadian instrument onboard the Swedish Odin satellite that has been operational since 2002 to the time of writing of this thesis. It uses the limb scattered sunlight approach, measuring wavelengths from 280 to 800 nm. The SASKTRAN radiative transfer model is used in combination with the obtained data to retrieve profiles of ozone number density with a spatial sampling of 1.0 km in altitude. This is a University of Saskatchewan instrument.

OMPS is a three-part instrument built by Ball Aerospace that is onboard the Suomi

National Polar-Orbiting Partnership satellite and has been operation since 2011. Its three parts are its nadir mapper, nadir profiler, and limb profiler. The data the limb profiler generates is the data that is used in the SOO data record. This part of the instrument operates similarly to OSIRIS, measuring limb scattered sunlight from 290 to 1000 nm. NASA uses a radiative transfer model to retrieve profiles of ozone number density, and distribute this data publicly. However, the atmospheric research group at the University of Saskatchewan also publishes a similar OMPS data product using the SASKTRAN radiative transfer model instead. The spatial sampling of this data is 1.0 km in altitude. In this thesis, we use the SASKTRAN version of the OMPS data.

The SOO data record is reported on a spatial grid of latitude and altitude (not longitude) and temporally in periods of 1 month. This type of data reporting is referred to by the atmospheric science community as monthly zonal means (MZM). We choose not to divide the data by longitude as well because the atmosphere is said to be "well-mixed" longitudinally, meaning that concentrations of species such as ozone remain relatively constant over longitude with the same altitude and latitude. So, we decide to average over all longitude in a latitude region.

To obtain the SOO data record, MZM data of ozone number density is first obtained for each of the three instruments and then these data are merged into one. Each instrument makes scans that ultimately result in ozone number density as a function of altitude at some latitude and longitude. To obtain MZM data, simply all the data that is collected in this fashion within a month that falls in a given latitude and altitude region is averaged. The regions we use in this thesis are 10° wide in latitude from [65° S - 55° S] to [55° N - 65° N] for a total of 13 latitude regions and 1 km high in altitude from 18.5 km to 50.5 km for a total of 48 altitude regions. In the following subsection, we discuss how the MZM data from the three instruments are merged into one data record.

2.1.1 Merging Method

Consider the three ozone number density MZM time series for each of the three instruments for an individual altitude-latitude region. One set of three time series for a particular altitude-latitude region is shown in Figure 2.1. To merge the three time series, biases between them

are first accounted for. Whereby bias we mean a consistent offset between one time series to another. For example, in Figure 2.1 it can be seen by the eye that in the overlap period for OMPS and OSIRIS the data of both instruments follow a similar trajectory, but the OMPS data is consistently larger than the OSIRIS data. The bias between the two time series is calculated by finding the average difference between them for each of the 12 months in the overlap period and then averaging these 12 averages. It is done this way instead of just calculating the average difference without regard for the months so that all months are weighted equally in the calculation of the bias. Once these are found for both SAGE II and OMPS relative to OSIRIS, then the SAGE II and OMPS time series are shifted by these values towards OSIRIS, removing the biases. Following from this each time series is "deseasonalized". This means to remove the yearly periodic signal in the data. This is done by calculating the average value of the time series for each of the 12 months of the year and subtracting these values from each instance of the corresponding month in the time series. The result is a time series with a mean of zero where each data point is positive if it is higher than the average for that month and negative if it is lower than the average. Then, the average value of the entire OSIRIS time series is added to the three individual time series. This brings the data back up to its usual height. Then, for months in the overlap periods where there is data from both instruments, the two values are averaged. Finally, we can consider what is left after this as one single time series. To calculate a "relative anomaly" from here, which is what the final SOO data record is reported in, we subtract and then divide each value by the mean of the entire time series. The final merged time series from the three time series shown in Figure 2.1 as an example is shown in Figure 2.2.

This process is, of course, carried out for every altitude-latitude region of the data, and the resulting merged time series make up the SOO data record. We show a heat map in Figure 2.3 to show the SOO data record in a particular latitude region for all altitudes. A full picture of the SOO data record would be given by 12 additional plots for each other latitude region. As for error estimation in these data, the SOO data record is reported with standard deviations. We do not explain how these are constructed in this thesis.

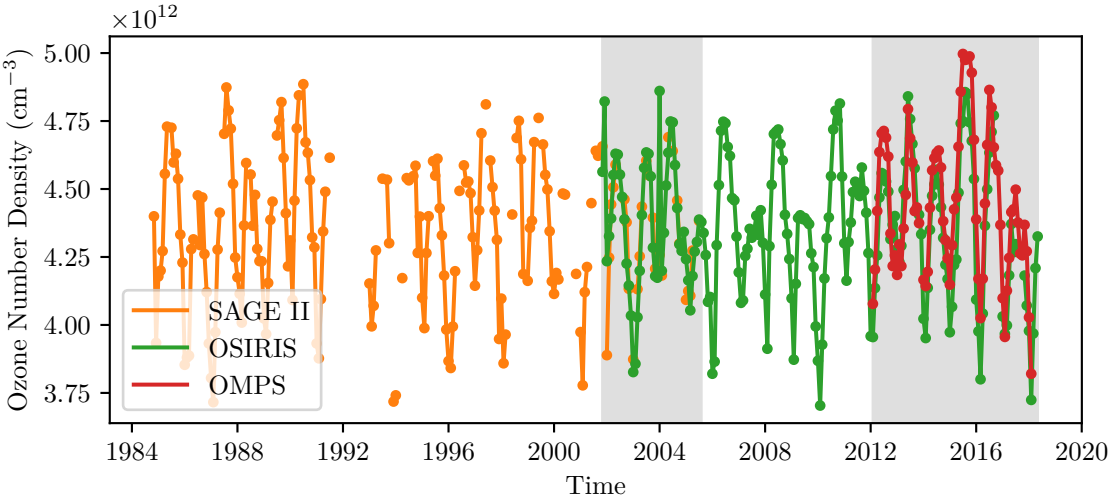


Figure 2.1: SAGE II, OSIRIS, and OMPS MZMs at an altitude of 24.5 km and latitude region of 5° to 15° N. Overlap periods between two instruments are in gray.

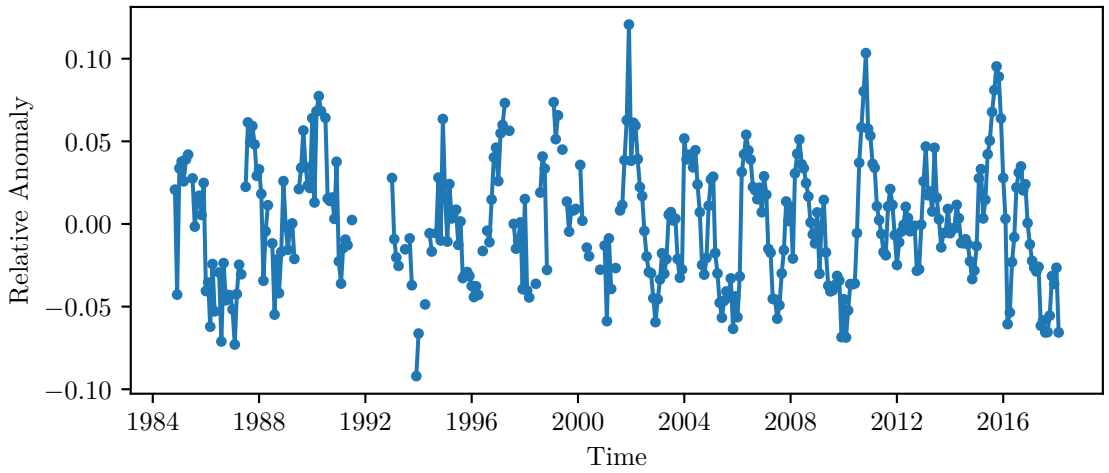


Figure 2.2: Merged MZM relative anomalies from the data in Figure 2.1.

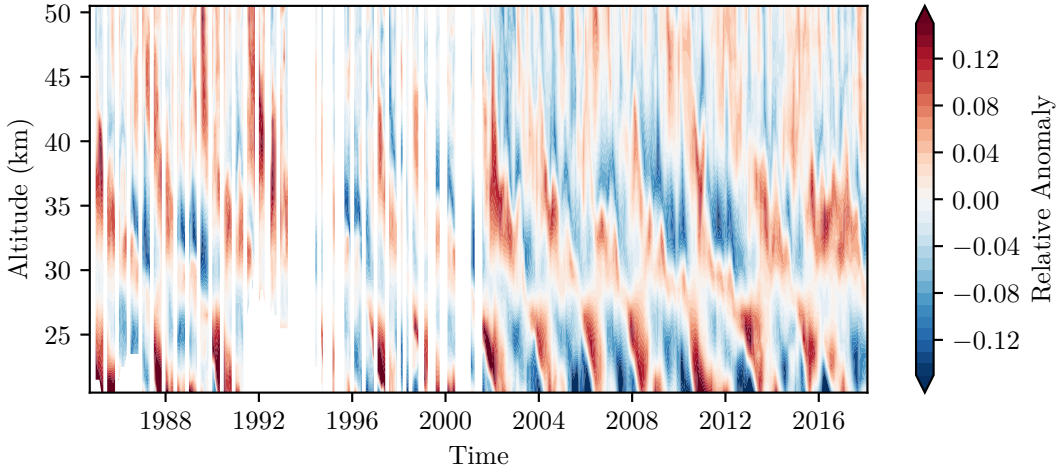


Figure 2.3: Merged SOO MZM relative anomalies for the latitude region 5° S to 5° N.

2.2 Ozone Influencing Phenomena

There are a number of atmospheric phenomena that stratospheric ozone is known to depend upon in one region or another. In this section, we give an overview of these phenomena and show times series "indexes" that describe them. These indexes are used in both Prais-Winsten and DLM procedures in this thesis. The indexes are obtained from standard sources about the phenomenon and we then scale them to have a mean of 0 and a standard deviation of 1.

These are phenomena that are, at best, weakly correlated with the yearly seasonal cycle. So, we find their signals in relative anomaly data records like SOO. The exact mathematical relationship between ozone in a given altitude-latitude region and these phenomena is of course not known. In light of this, what we do for the MLR and DLM statistical models is assume that the relationship is linear. So, we say that an ozone time series can be approximated as a linear combination of the indexes that are presented in this section as:

$$OZONE(t) = a_1 Phenomenon1(t) + a_2 Phenomenon2(t) + ... \tag{2.1}$$

where $a_1, a_2, ...$ are constants. This is discussed further in Section 2.3. The following sections

discuss the ozone influencing phenomena and the indexes we use to describe them.

2.2.1 The Quasi-Biennial Oscillation

The Quasi-Biennial Oscillation (QBO) is a quasiperiodic oscillation of east moving and west moving wind in the tropical stratospheric region and is known to influence ozone in both the tropical and subtropical stratosphere. The mean period of the QBO is 28 to 29 months. To create an index for the QBO we use wind observations collected from radiosonde stations near the equator, primarily from Singapore. The data is reported as monthly means produced from daily observations on pressure altitude surfaces of 70, 50, 40, 30, 20, 15, and 10 hPa. It is known that this data can be considered representative of the stratospheric winds in a belt around the entire globe near the equator.

Rather than assuming ozone is a linear combination of the data at each of these altitudes, or only one of these altitudes perhaps, a principal component analysis is performed on the data for the 7 altitudes to obtain two uncorrelated time series indexes. Each of the 7 altitude time series can be represented fairly accurately by a linear combination of these two time series. Therefore, in theory, writing ozone as a linear combination of these for the statistical models is just as good as writing it as a linear combination of all 7. But, it is in fact probably better considering that more terms in the linear combination can cause problems for the statistical models.

We call these two principal components QBOA and QBOB. The generated QBOA and QBOB time series are shown in Figure 2.4, having been scaled to a mean of 0 and a standard deviation of 1.

2.2.2 The Solar Cycle

The solar cycle is an 11-year periodic change in the solar activity of the Sun that is known to influence stratospheric ozone. For the index to describe it, we use what is called the F10.7 index. This is a measurement of flux values emitted from the sun at a wavelength of 10.7 cm. This data has been measured daily in Canada in either Ottawa or Pentiction since 1947 and is reported as monthly means. We call this index the SOLAR index in this thesis and is

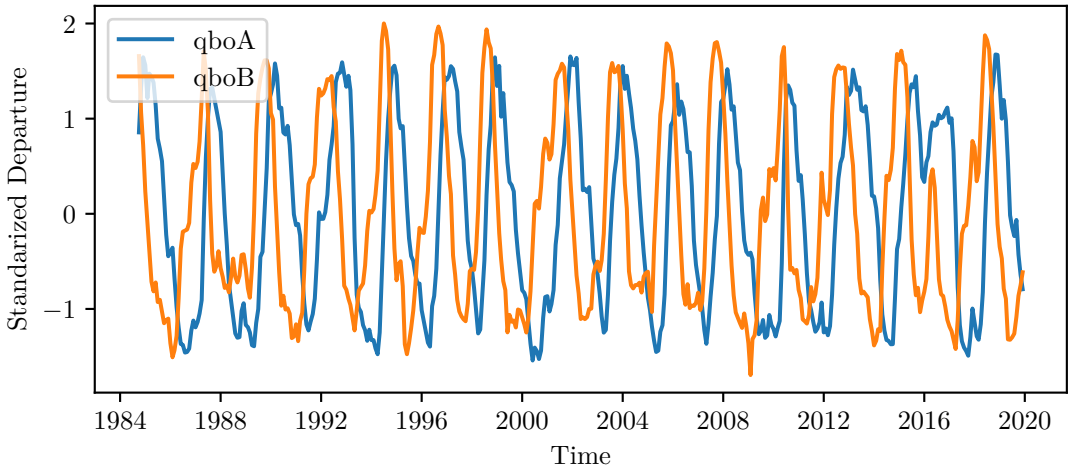


Figure 2.4: The QBOA and QBOB Indexes (principal components of Singapore winds).

shown in Figure 2.5 scaled to a mean of 0 and a standard deviation of 1.

2.2.3 The El Niño-Southern Oscillation

The El Niño-Southern Oscillation (ENSO) is associated with warm water that develops in the east-central pacific ocean that is known to be coupled to many atmospheric phenomena, including ozone concentration. For the index to describe it, we use what is called the Multivariate ENSO Index, which is compiled based on six observed variables over the tropical pacific (sea-level pressure, zonal and meridional components of the surface wind, sea surface temperature, surface air temperature, and total cloudiness fraction of the sky) (Wolter and Timlin, 2011). This time series is shown in Figure 2.6, scaled to a mean of 0 and a standard deviation of 1.

2.2.4 Aerosols

Aerosols are known to interact with ozone and affect their concentration. After the large eruption of Mt. Pinatubo in 1992, a significant response from ozone is seen in many altitude-latitude regions of SAGE II data. For the index, we use the aerosol optical depth at 550

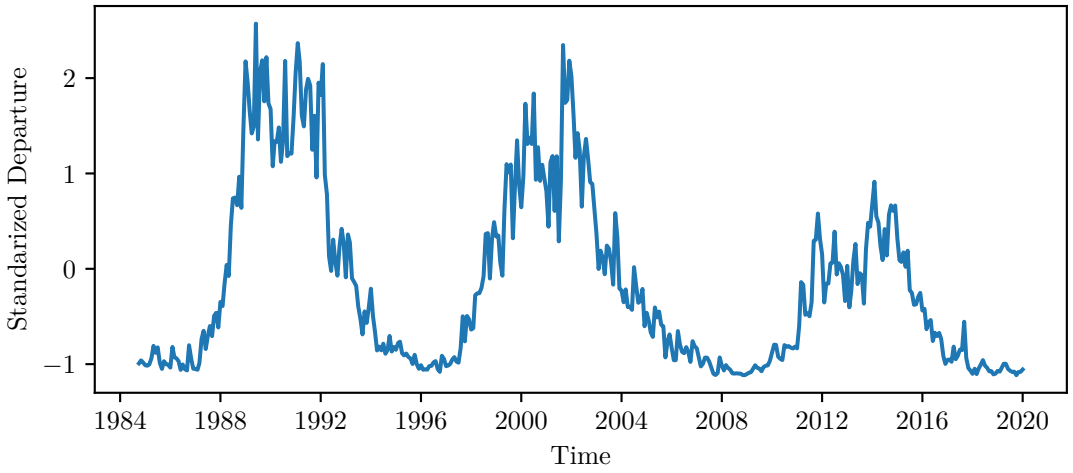


Figure 2.5: The SOLAR Index.

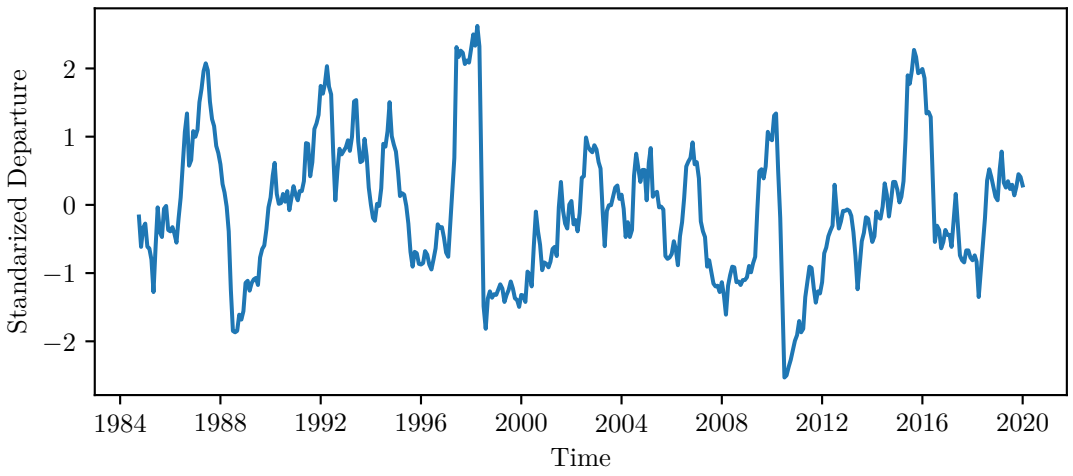


Figure 2.6: The ENSO Index.

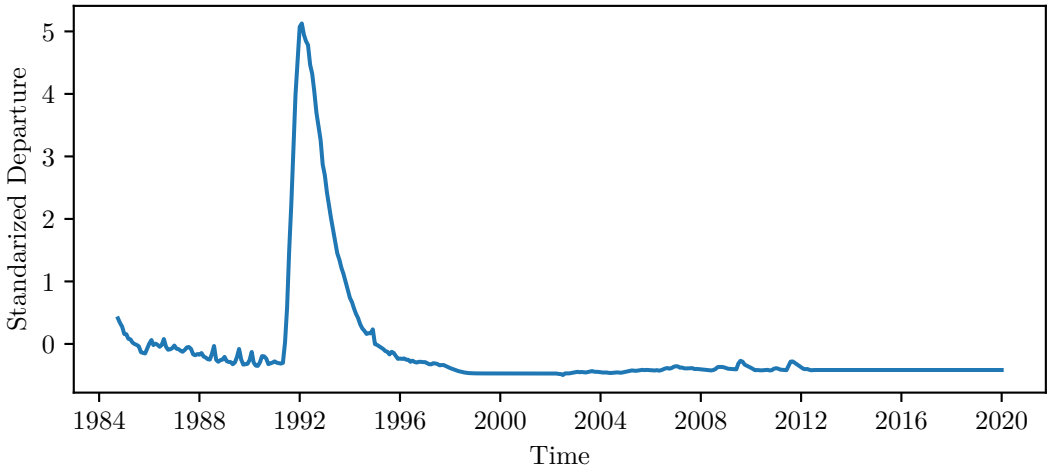


Figure 2.7: The AOD Index.

nm in the altitude region of 15 to 35 km. This data is published by NASA's Goddard Institute for Space Studies. The data is given as a function of latitude and altitude. So, if desired a separate time series index can be used in the linear combination for ozone for each corresponding altitude-latitude region of SOO relative anomalies. But, stratospheric aerosol only really affects ozone in the lower altitudes and there is not much variation in the shape of these time series with latitude. So, we use a global average of the optical depth data referred to as our index for the statistical models for all the altitude-latitude regions of SOO. This time series, scaled to a mean of 0 and a standard deviation of 1, is shown in Figure 2.7. A large spike in the time series around the year 1992 corresponding to the eruption of Mt. Pinatubo can be seen. We refer to this index as the aerosol optical depth (AOD) index.

2.2.5 Linear Increases and Decreases

Since the beginning of the SOO data record, there have been trends in ozone concentration in the positive and negative direction. This has been found to be due to the changing amount of chlorine in the atmosphere (mostly man-made). In the high altitudes and high to middle latitudes for most ozone data records, including SOO, it is found that ozone typically decreases until about the year 1997 and then increases at a slower rate thereafter. We make

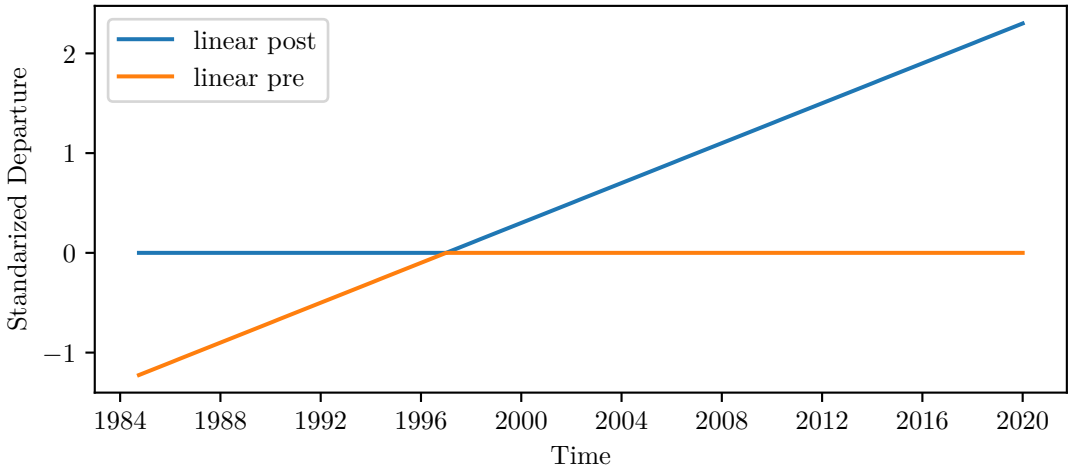


Figure 2.8: The LINEAR POST and LINEAR PRE Indexes.

two time series indexes, like the ones we have presented above, to represent this. They are shown in Figure 2.8. This index is used in the Prais-Winsten estimation but not the DLM procedure. The blue line in the figure is 0 before 1997 and linearly increasing after and the orange line is just the opposite. We refer to these indexes as LINEAR POST and LINEAR PRE respectively.

2.3 Modelling Ozone

As stated in Section 2.2, we can assume ozone is a linear combination of all the atmospheric phenomena listed in Section 2.2, including the LINEAR PRE and LINEAR POST indexes that we defined. More specifically, for the MLR model, we approximate an ozone time series as,

$$\begin{aligned} OZONE(t) = a_1 QBOA(t) + a_2 QBOB(t) + a_3 SOLAR(t) + a_4 ENSO(t) + a_5 AOD(t) \\ + a_6 LINEARPRE(t) + a_7 LINEARPOST(t) + a_8, \end{aligned} \tag{2.2}$$

where $OZONE(t)$ is the ozone approximation, $QBOA(t), QBOB(t), ..., LINEARPOST(t)$ are the time series indexes presented in the last section in Figures 2.4 through 2.8, and

$a_1, a_2, .., a_7$ are the coefficients multiplying each of them. A constant a_8 is also included because when finding the coefficients for the MLR model there is often an offset that needs to be corrected for.

In the next section, we cover the theory of MLR and techniques to estimate these coefficients, including the Prais-Winsten estimation procedure. Until recently the Prais-Winsten estimation procedure represented the state-of-the-art for ozone trend analysis. The section goes into rigorous detail, but simply, with the most common technique, the coefficients are found such that the sum of the squared differences between the approximation $OZONE(t)$ and the actual observed time series at each time t is a minimum. The Prais-Winsten estimation is a slight modification to this principle, where an optimization is done considering that the observed time series may be autocorrelated. This leads to similar, but theoretically more optimized, estimate and is discussed later in the section after the basic least squares theory is presented.

2.4 Multiple Linear Regression

Regression analysis is a commonly used statistical technique for investigating the relationship between variables. Multiple linear regression (MLR) allows for the expression of the relationship between a variable of interest as a linear combination of a set of variables that the variable of interest is known to depend upon. The variable of interest is often called the response variable and the variables it depends on are often called predictors, regressors, or proxies. For the ozone topic of this thesis, we have set up ozone as the response variable and the ozone influencing phenomena, LINEAR PRE, and LINEAR POST indexes as the predictors. In Section 2.4.1 we cover the background theory for the MLR model using what is called the ordinary least squares (OLS) estimation technique. This paves the way for Section 2.4.2, where we cover the background theory for what is called the generalized least squares (GLS) estimation technique and Section 2.4.3 on the Prais-Winsten estimation. Prais-Winsten estimation is a procedure that is fundamentally a GLS estimation technique that optimizes a given MLR model and improves the error analysis. In Section 2.4.5 we show how to quantify statistical uncertainty in the MLR model, and in Section 2.4.6 we summarize

the Prais-Winsten estimation algorithm for easy reference.

2.4.1 Basic Theory

If there are k predictors with known values $x_1, x_2, ..., x_k$ then the MLR model treats the response variable Y as a random variable and assumes that its expected value is a linear combination of the predictors. So we have,

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k, \tag{2.3}$$

where $\beta_1, ..., \beta_k$ are the linear coefficients of the predictors and β_0 is an added constant term that is necessary for almost all problems in practice. These coefficients are referred to as the regression coefficients.

The variance of the random variable Y can be defined as σ^2. So, if we define a random variable e to have expectation 0 and variance σ^2 (to denote this we simply write $e \sim [0, \sigma^2]$) then Y can be written as,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + e, \qquad e \sim [0, \sigma^2]. \tag{2.4}$$

Consider now that the values of the predictors change with time, and that there are n known sets of these values at various times. The predictor variable sets will be defined as,

$$(x_{11}, x_{21}, ..., x_{k1}), (x_{12}, x_{22}, ..., x_{k2}), ..., (x_{1n}, x_{2n}, ..., x_{kn}). \tag{2.5}$$

Then, the random variables $Y_1, Y_2, ..., Y_n$, which we define to represent the response variable for each predictor variable set, are then specified as,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + e_i, \qquad e_i \sim [0, \sigma^2] \qquad i = 1, 2, ..., n, \tag{2.6}$$

where the e_i are independent random variables with mean 0 and variance σ^2. This is now the typical form that the MLR model is presented as. Furthermore, Equation 2.6 can be written more compactly in matrix form as,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \qquad \mathbf{e} \sim [0, \sigma^2\mathbf{I}], \tag{2.7}$$

where $\mathbf{Y}$ is a n x 1 vector of the random variables $Y_1, Y_2, ..., Y_n$, $\mathbf{X}$ is an n x $(k+1)$ matrix of each of the predictor variable sets with a first column of 1's, $\boldsymbol{\beta}$ is a $(k+1)$ x 1 vector of the regression coefficients, and $\mathbf{e}$ is an n x 1 vector of the random variables $e_1, e_2, ..., e_n$.

Lastly, it should be noted that for the MLR model, there is no reason that we need to assume all the e_i are the same random variables. But, we will work under this assumption for now and see later via the Gauss-Markov theorem that the least squares estimate we use for the $\boldsymbol{\beta}$ coefficients is optimal in a sense under this assumption. Furthermore, we will use our understanding of this to find a similar optimal estimate when this assumption does not hold. This is the main idea behind the GLS estimation technique discussed later in Section 2.4.2. But, note that under this assumption the covariance matrix of $\mathbf{e}$ and $\mathbf{Y}$ is $\sigma^2\mathbf{I}$ as shown in Equation 2.7. This follows clearly from the fact that the e_i are independent, meaning that,

$$\mathrm{E}[e_k e_j] = \begin{cases} \sigma^2 & k = j \\ 0 & k \neq j \end{cases}. \tag{2.8}$$

The interpretation of this covariance matrix is that the variance of the errors (or just as well, the variance of the Y_i's) is the same for all indexes 1 to n and that there is no correlation from one index to another.

2.4.1.1 Determining Model Parameters

The model of the response variable is fully specified by the model parameters $\beta_0, ..., \beta_k$, and σ^2. If these parameters are not known then they can be estimated from measurements. Consider that n measurements of the response variable are made that we denote by the values $y_1, y_2, ..., y_n$, and that they correspond to each predictor variable set. Specifically, the following tuples are observed:

$$(y_1, x_{11}, x_{21}, ..., x_{k1}), (y_2, x_{12}, x_{22}, ..., x_{k2}), ..., (y_n, x_{1n}, x_{2n}, ..., x_{kn}). \tag{2.9}$$

As a side note, notice that this is the case we have for the ozone problem of this thesis. We have monthly ozone data as the measurements, y_i, and monthly ozone influencing phenomena as the predictor variables, x_{ij}.

Now, owing to the form of the MLR model that has been given, the following n equations can be written:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + ... + \beta_k x_{k1} + \epsilon_1 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + ... + \beta_k x_{kn} + \epsilon_n, \end{aligned} \tag{2.10}$$

where the ϵ_i's are numbers that make these equations hold given the tuples of 2.9 and whatever regression coefficients are used. The method most widely used to estimate the regression coefficient model parameters $\beta_0, ..., \beta_k$ is to choose them so that the sum of the squares of the ϵ_i's in these equations are minimized (i.e. $\epsilon_1^2 + \epsilon_2^2 + ... + \epsilon_n^2$ is minimized). This is called the method of least squares or OLS.

This minimization problem can be easier solved if the equations of 2.10 are written in matrix form. So, we write these equations as,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.11}$$

where $\mathbf{y}$ is an n x 1 vector with elements $y_1, ..., y_n$, $\boldsymbol{\epsilon}$ is an n x 1 vector with elements $\epsilon_1, ..., \epsilon_n$, and $\mathbf{X}$ and $\boldsymbol{\beta}$ have been defined previously. The sum of the squares of the ϵ_i's will now be written and minimized with respect to $\boldsymbol{\beta}$. This sum, $S(\boldsymbol{\beta})$, is given by,

$$S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.12}$$

and the solution for $\boldsymbol{\beta}$ from setting $dS(\boldsymbol{\beta})/d\boldsymbol{\beta} = 0$ is (see Appendix A for these details)

$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{2.13}$$

This is the OLS estimate of the regression coefficients for an MLR model given the observations specified by 2.9. We define this quantity as $\boldsymbol{\beta}^*$.

To estimate the final MLR model parameter σ^2, Equation 2.11 will be rewritten with the OLS estimate $\boldsymbol{\beta}^*$ as,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}^*, \tag{2.14}$$

where the elements of the $\boldsymbol{\epsilon}^*$ vector in this equation are called the residuals. The residuals are the differences between the observed values $\mathbf{y}$ and the estimated mean values of $\mathbf{Y}$, $\mathbf{X}\boldsymbol{\beta}^*$. The residuals $\boldsymbol{\epsilon}^*$ are often confused with the errors $\mathbf{e}$ by beginners, so we make sure to make the distinction here. Now, an unbiased estimate of σ^2 can be obtained from these residuals and is given by (see Appendix B for proof that this is unbiased and see Appendix C for the definition of bias of an estimator),

$$\sigma^{2^*} = \frac{\boldsymbol{\epsilon}^{*^\mathbf{T}}\boldsymbol{\epsilon}^*}{n-(k+1)}. \tag{2.15}$$

2.4.1.2 The Ordinary Least Squares Estimator

By substituting $\mathbf{Y}$ in place of $\mathbf{y}$ in Equation 2.13, the OLS regression coefficient estimator, which we define as $\hat{\boldsymbol{\beta}}$, is obtained as,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{2.16}$$

The covariance matrix and expected value of $\hat{\boldsymbol{\beta}}$ will now be computed. Taking the expected value of Equation 2.16 it is found that

$$\mathrm{E}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{E}[\mathbf{Y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}. \tag{2.17}$$

Therefore, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. Taking the covariance of Equation 2.16 (using the identity shown in Appendix D) gives

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Cov}[\mathbf{Y}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}. \tag{2.18}$$

Now, since $\mathrm{Cov}[\mathbf{Y}] = \sigma^2\mathbf{I}$ under the assumption discussed prior, this can be simplified to

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \tag{2.19}$$

2.4.2 Generalized Least Squares Estimation

With the assumption discussed that the expected value and covariance matrix of the errors is a zero vector and $\sigma^2\mathbf{I}$ respectively, the Gauss-Markov theorem states that the OLS estimator of the regression coefficients given by Equation 2.13 is the unbiased estimator with the minimum variance out of all other possible linear in $\mathbf{Y}$ estimators (see Appendix E for this proof). So, the estimator is optimal in this sense, and this is typically called the best linear unbiased estimator (BLUE). If this assumption does not hold however, the OLS estimate can be altered so that a BLUE is still obtained. This is called the Generalized Least Squares (GLS) estimate. This is only possible however if another covariance structure of the errors is specified. The theory behind this GLS estimate is that in knowing the true covariance structure a transformation can be made to the MLR model equation such that the error term becomes a constant multiplied by the identity matrix yet again. This technique is described in this section.

To determine whether data warrants the use of the GLS over the OLS, we can look at the residuals of the MLR model using the OLS estimate. Essentially, if the assumption of the errors e_i all being the same and uncorrelated is good, then the residuals should closely approximate draws from the probability distribution of the random variable e_i. If this is not the case then there is something about the MLR model that does not accurately represent the given data. Often, but not always, this can just be the assumption about the errors e_i being all the same. If this is the case then this is the indication that the GLS estimation is better for the given data (assuming that some reasonable estimation about the true covariance structure of the errors can be made). There are more rigorous mathematical ways to assess the residuals for this, but often it is just recommended to look at scatter plots of the residuals versus its index and see if the plot resembles draws from the assumed e_i or not. If it does not, then a person can usually see in what way it does not. For example, if we look at the residuals using OLS estimation for SOO ozone data at various altitude-latitude regions, we often see that the values at one time are likely to be similar to the values at the time previous. This indicates that the ozone data is probably serially correlated.

In Section 2.4.2.1 the GLS technique of transforming the MLR model equation is shown.

In Section 2.4.2.2 the GLS estimate of the regression coefficients is given, and in Section 2.4.2.3 a general method for estimating the covariance structure of the errors so that the GLS can be used in practice is given.

2.4.2.1 The Transformed Model

Instead of assuming the covariance matrix of $\mathbf{e}$ is $\sigma^2\mathbf{I}$, like in the MLR model we have already presented, a general covariance matrix in the form of $a^2\mathbf{V}$ is assumed, where $\mathbf{V}$ is some symmetric positive definite matrix and a^2 is a positive constant. So, the MLR model equation becomes

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \qquad \mathbf{e} \sim [0, a^2\mathbf{V}]. \tag{2.20}$$

Since $\mathbf{V}$ is a symmetric positive definite matrix, a Cholesky decomposition can be performed on it such that $\mathbf{V} = \mathbf{K}\mathbf{K}^T$, where the decomposition is unique (i.e. there exists only one matrix $\mathbf{K}$ that satisfies this). For convenience we will define $\mathbf{G}$ as $\mathbf{K}^{-1}$. Now, the primary property that allows for the GLS technique to work in finding the BLUE is that it turns out that the covariance matrix of $\mathbf{Ge}$ is $a^2\mathbf{I}$. So, the MLR model equation can be right multiplied by the matrix $\mathbf{G}$,

$$\mathbf{GY} = \mathbf{GX}\boldsymbol{\beta} + \mathbf{Ge}, \qquad \mathbf{e} \sim [0, a^2\mathbf{V}], \tag{2.21}$$

so that the final term has a covariance structure of $a^2\mathbf{I}$. This is in principal identical to the MLR model we had before. We can just consider $\mathbf{GY}$ to be $\mathbf{Y}$ from before, $\mathbf{GX}$ to be $\mathbf{X}$ from before, and $\mathbf{Ge}$ to be $\mathbf{e}$ from before and it is the same model equation. Then importantly, since the $\mathbf{Ge}$ term has covariance $a^2\mathbf{I}$, we know that the least squares estimate of the regression coefficients with this transformed model results in a BLUE. Lastly, we give the simple proof that the covariance matrix of $\mathbf{Ge}$ is $a^2\mathbf{I}$,

$$
\begin{aligned}
\text{Cov}[\mathbf{Ge}] &= \text{E}[\mathbf{Ge}(\mathbf{Ge})^T] - \text{E}[\mathbf{Ge}]\text{E}[\mathbf{Ge}]^T \\
&= \text{E}[\mathbf{Ge}(\mathbf{Ge})^T] \\
&= \mathbf{G}\text{E}[\mathbf{ee}^T]\mathbf{G}^T \\
&= a^2\mathbf{GVG}^T \\
&= a^2\mathbf{GG}^{-1}\mathbf{G}^{-1^T}\mathbf{G}^T \\
&= a^2\mathbf{I}.
\end{aligned}
\tag{2.22}
$$

2.4.2.2 The Regression Coefficient Estimates

To find the GLS estimate of the regression coefficients we can use the same technique as for the OLS estimate where we wrote n equations for each of the data observations defined by 2.9. We can show this as,

$$
\mathbf{Gy} = \mathbf{GX}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.23}
$$

Again, minimizing the sum of the squares of the elements of $\boldsymbol{\epsilon}$ results in the GLS BLUE. Alternatively, we could easily just use our knowledge of the transformed MLR model equation and the OLS estimate obtained prior to quickly find the GLS estimate (see Appendix A for these details). The GLS estimate is given as,

$$
\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}, \tag{2.24}
$$

and the corresponding estimator is

$$
\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}, \tag{2.25}
$$

using the same notation that we used for OLS of $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}$.

The expected value and covariance matrix of this estimator will now be computed. The expected value is calculated as,

$$
\text{E}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\text{E}[\mathbf{Y}] = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}. \tag{2.26}
$$

Therefore, $\hat{\boldsymbol{\beta}}$ is unbiased (as we already know should be the case from the Gauss-Markov theorem). By taking the covariance, the covariance matrix of this estimator is found to be,

$$\text{Cov}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\text{Cov}[\mathbf{Y}]\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{2.27}$$

$$= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}a^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{2.28}$$

$$= a^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{2.29}$$

2.4.2.3 Estimating the Matrix V

To perform a GLS estimation, it is required that the matrix $\mathbf{V}$ is known. Unfortunately, in most applications it is not known, nor is it simple to determine what it is. However, an estimate of $\mathbf{V}$ can be obtained by a method known as Feasible Generalized Least Squares (FGLS). The idea behind FGLS is to first perform the OLS estimate so that the resulting residuals can be used to estimate $\mathbf{V}$. Then, this estimated $\mathbf{V}$ is used to perform a GLS estimation, to hopefully obtain a better estimate than OLS. To improve this estimate further, residuals from the GLS estimation are used to update the estimate of $\mathbf{V}$ so that another GLS estimation can be performed. This is then done iteratively in the method known as FGLS until sequential $\mathbf{V}$'s vary by less than some tolerance. The Prais-Winsten estimation procedure that we use for stratospheric ozone is a special case of FGLS, the Prais-Winsten estimation is the content of the next section. Let us define the general structure of $a^2\mathbf{V}$ for the MLR model with the GLS technique as,

$$a^2\mathbf{V} = \begin{bmatrix} \text{Var}[e_1] & \text{Cov}[e_1, e_2] & \dots & \text{Cov}[e_1, e_n] \\ \text{Cov}[e_1, e_2] & \text{Var}[e_2] & \dots & \text{Cov}[e_2, e_n] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[e_1, e_n] & \text{Cov}[e_2, e_n] & \dots & \text{Var}[e_n] \end{bmatrix}, \tag{2.30}$$

where $e_1, ..., e_n$ are the elements of $\mathbf{e}$.

2.4.3 Prais-Winsten Estimation

The Prais-Winsten estimation procedure can be used to estimate $\mathbf{V}$ for the GLS estimation when it is assumed that the only covariance structure in the errors $\mathbf{e}$ is autocorrelation (Prais and Winsten, 1954). More specifically, when this autocorrelation is assumed to be described by an autoregressive model of first order (denoted as AR(1)).

The AR(1) model for a data sequence $z_1, ..., z_n$ modelled by the random variables $Z_1, ..., Z_n$ is given as,

$$Z_i = c + \rho Z_{i-1} + \xi_i, \qquad \xi_i \sim [0, \sigma_e^2] \qquad i = 2, 3, ..., n, \tag{2.31}$$

where ρ is the main parameter of the model that correlates the previous index to the current index, c is a constant, and $\xi_2, ...\xi_n$ are independent random variables with mean 0 and variance σ_e^2. This is called an autoregressive model because it actually has the structure of an MLR model. It effectively is an MLR model with Z_{t-1} as the single regressor.

We will now calculate the theoretical $a^2\mathbf{V}$ matrix given in Equation 2.30 when the error random variables $e_1, ..., e_n$ of the MLR model are assumed to be described by an AR(1) model such that,

$$e_i = \rho e_{i-1} + \xi_i, \qquad \xi_i \sim [0, \sigma_e^2] \qquad i = 2, 3, ..., n. \tag{2.32}$$

We drop the constant c for this because we have $\mathrm{E}[e_i] = 0$ for all i in the MLR model. To calculate the diagonal elements of $a^2\mathbf{V}$, we take the variance of Equation 2.32,

$$\mathrm{Var}[e_i] = \mathrm{Var}[\rho e_{i-1}] + \mathrm{Var}[\xi_i] \tag{2.33}$$

$$= \rho^2 Var(e_{i-1}) + \sigma_e^2. \tag{2.34}$$

Since the assumption of equal variances of all the errors is being made (i.e. $Var(e_i) = Var(e_{i-1})$) we have,

$$Var(e_i) = \frac{\sigma_e^2}{1 - \rho^2}. \tag{2.35}$$

For all the off-diagonal elements of $a^2\mathbf{V}$, the covariance of e_i and e_{i-s} will be found where s represents the difference in the index between the two random variables. First, notice that the following equation can be written:

$$e_{i-1} = \rho e_{i-2} + \xi_{i-1}. \tag{2.36}$$

Using this equation and Equation 2.32, it can be written that

$$e_i = \rho^2 e_{i-2} + \rho\xi_{i-1} + \xi_i. \tag{2.37}$$

If this is done iteratively (i.e. writing e_i in terms of e_{i-3}, then e_{i-4}, and so on) it can be found that

$$e_i = \rho^s e_{i-s} + \sum_{j=1}^{s} \rho^{s-j}\xi_{i-(s-j)} = \rho^s e_{i-s} + \sum_{j=0}^{s-1} \rho^j \xi_{i-j}. \tag{2.38}$$

Now, using the definition of the covariance of random variables we have,

$$\begin{aligned}
\mathrm{Cov}[e_i, e_{i-s}] &= \mathrm{E}[(e_i - \mathrm{E}[e_i])(e_{i-s} - \mathrm{E}[e_{i-s}])] && (2.39)\\
&= \mathrm{E}[e_i e_{i-s}] && (2.40)\\
&= \mathrm{E}[(\rho^s e_{i-s} + \sum_{j=0}^{s-1} \rho^j \xi_{i-j}) e_{i-s}] && (2.41)\\
&= \rho^s \mathrm{E}[e_{i-s} e_{i-s}] + \sum_{j=0}^{s-1} \rho^j \mathrm{E}[\xi_{i-j} e_{i-s}] && (2.42)\\
&= \rho^s \mathrm{E}[e_{i-s} e_{i-s}] && (2.43)\\
&= \rho^s Var(e_{i-s}) && (2.44)\\
&= \rho^s \frac{\sigma_e^2}{1-\rho^2}. && (2.45)
\end{aligned}$$

The fourth step requires noticing that $\mathrm{E}[\xi_a e_b] = 0$ for $a > b$. So, by putting these results into the general matrix specified by Equation 2.30, we have $a^2\mathbf{V}$ under this AR(1) autocorrelation assumption with $a^2 = \sigma_e^2$ and $\mathbf{V}$ as,

$$\mathbf{V} = \begin{bmatrix} \frac{1}{1-\rho^2} & \frac{\rho}{1-\rho^2} & \cdots & \frac{\rho^{n-1}}{1-\rho^2} \\ \frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & \cdots & \frac{\rho^{n-2}}{1-\rho^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\rho^{n-1}}{1-\rho^2} & \frac{\rho^{n-2}}{1-\rho^2} & \cdots & \frac{1}{1-\rho^2} \end{bmatrix}. \tag{2.46}$$

Furthermore, it can be verified that $\mathbf{V^{-1}}$ is (see Appendix F for details)

$$\mathbf{V}^{-1} = \begin{bmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix}, \tag{2.47}$$

and that the matrix $\mathbf{G}$ from the Cholesky decomposition is

$$\mathbf{G} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 & 0 \\ -\rho & 1 & 0 & \dots & 0 & 0 \\ 0 & -\rho & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix}. \tag{2.48}$$

2.4.3.1 Prais-Winsten Estimation with Missing Data

In the previous section, $\mathbf{V}$ was explicitly calculated from its general form, its inverse was then found, and then the Cholesky decomposition of this inverse was done to find the matrix $\mathbf{G}$. Even though $\mathbf{G}$ is not necessarily needed to estimate $\hat{\boldsymbol{\beta}}$ or $\text{Cov}[\hat{\boldsymbol{\beta}}]$ for the GLS model with $\mathbf{V}^{-1}$ given, it is informative from a theoretical perspective since it is the matrix applied to transform the MLR model equation. Alternatively to the approach in the previous section, a matrix $\mathbf{M}$ could have been found such that $\text{Cov}[\mathbf{Me}] = a^2\mathbf{I}$. Then, considering the proof carried out in Equation 2.22, it can be concluded that such a matrix must equal the matrix $\mathbf{G}$. If $\mathbf{G}$ can be found in this way then $\mathbf{V}^{-1}$ is automatically known by a matrix multiplication

(i.e. $\mathbf{V^{-1}} = \mathbf{G}^T\mathbf{G}$). The problem of correcting for missing observations is looked at with this alternative approach.

First, we will give the verification that the $\mathbf{G}$ we have already seen in Equation 2.48 without considering missing data results in $\text{Cov}[\mathbf{Ge}] = a^2\mathbf{I}$. Then we will do the same for another $\mathbf{G}$, given by reference (Savin and White, 1978), that accounts for missing data. Without missing data, we start by evaluating $\mathbf{Ge}$ as,

$$\mathbf{Ge} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 & 0 \\ -\rho & 1 & 0 & \dots & 0 & 0 \\ 0 & -\rho & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_{n-1} \\ e_n \end{bmatrix} = \begin{bmatrix} \sqrt{1-\rho^2}e_1 \\ -\rho e_1 + e_2 \\ -\rho e_2 + e_3 \\ \vdots \\ \vdots \\ -\rho e_{n-1} + e_n \end{bmatrix} = \begin{bmatrix} \sqrt{1-\rho^2}e_1 \\ \xi_2 \\ \xi_3 \\ \vdots \\ \vdots \\ \xi_n \end{bmatrix}. \quad (2.49)$$

The last step is done with Equation 2.32. The covariance matrix of this vector is

$$\text{Cov}[\mathbf{Ge}] = \text{E}[\mathbf{Ge}(\mathbf{Ge})^T] - \text{E}[\mathbf{Ge}]\text{E}[\mathbf{Ge}]^T \quad (2.50)$$

$$= \text{E}[\mathbf{Ge}(\mathbf{Ge})^T]. \quad (2.51)$$

Now, since $\text{E}[\xi_t\xi_s] = 0$ for any non-equal t and s and $\text{E}[\xi_t e_1] = 0$ for any $t > 1$, all off-diagonal elements of $\text{Cov}[\mathbf{Ge}]$ are zero. For the diagonal elements, we know that $\text{E}[\xi_t^2] = \sigma_e^2$ for any t, and therefore we know all the diagonal elements except for the first are σ_e^2. For the first element, the expectation can be carried out as,

$$\text{E}[(\sqrt{1-\rho^2}e_1)^2] = (1-\rho^2)\text{E}[e_1^2] = (1-\rho^2)\frac{\sigma_e^2}{(1-\rho^2)} = \sigma_e^2, \quad (2.52)$$

noticing that $\text{E}[e_1^2] = Var(e_1)$. So, the covariance matrix $\text{Cov}[\mathbf{Ge}]$ with $\mathbf{G}$ specified in Equation 2.48 is indeed $a^2\mathbf{I}$ where $a^2 = \sigma_e^2$. Therefore, we can conclude that this is the correct matrix $\mathbf{G}$ that should be used in the GLS theory for an AR(1) covariance structure in the errors, in the absence of missing observations.

Now, we will give the matrix $\mathbf{G}$ for missing observations and then verify that it gives

$\text{Cov}[\mathbf{Ge}] = \sigma_e^2\mathbf{I}$, like we have just done above. When there are m missing observations following some index s of the data, the matrix $\mathbf{G}$ is modified to be,

$$\mathbf{G} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ -\rho & 1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & -g\rho^{m+1} & g & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & -\rho & 1 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix} \tag{2.53}$$

where g is given as,

$$g^2 = (\sum_{r=0}^{m} \rho^{2r})^{-1} = \frac{1-\rho^2}{1-\rho^{2(m+1)}}. \tag{2.54}$$

Now, we see that evaluating $\mathbf{Ge}$ gives,

$$\mathbf{Ge} = \begin{bmatrix} \sqrt{1-\rho^2}e_1 \\ -\rho e_1 + e_2 \\ \vdots \\ -\rho e_{s-1} + e_s \\ -g\rho^{m+1}e_s + ge_{s+m+1} \\ -\rho e_{s+m+1} + e_{s+m+2} \\ \vdots \\ \vdots \\ -\rho e_{n-1} + e_n \end{bmatrix}. \tag{2.55}$$

To simplify the $-g\rho^{m+1}e_s + ge_{s+m+1}$ expression that shows up, we can see from Equation 2.38 that we can write,

$$e_{s+m+1} = \rho^{m+1}e_s + \sum_{j=0}^{m} \rho^j \xi_{s+m+1-j}, \tag{2.56}$$

and therefore we can also write,

$$-g\rho^{m+1}e_s + ge_{s+m+1} = g\sum_{j=0}^{m}\rho^j\xi_{s+m+1-j}. \tag{2.57}$$

So, we simplify **Ge** further with this result as,

$$\mathbf{Ge} = \begin{bmatrix} \sqrt{1-\rho^2}e_1 \\ -\rho e_1 + e_2 \\ \vdots \\ -\rho e_{s-1} + e_s \\ -g\rho^{m+1}e_s + ge_{s+m+1} \\ -\rho e_{s+m+1} + e_{s+m+2} \\ \vdots \\ \vdots \\ -\rho e_{n-1} + e_n \end{bmatrix} = \begin{bmatrix} \sqrt{1-\rho^2}e_1 \\ \xi_2 \\ \vdots \\ \xi_s \\ -g\sum_{j=0}^{m}\rho^j\xi_{s+m+1-j} \\ \xi_{s+m+2} \\ \vdots \\ \vdots \\ \xi_n \end{bmatrix}. \tag{2.58}$$

Now, like before we have $\text{Cov}[\mathbf{Ge}] = \text{E}[\mathbf{Ge}(\mathbf{Ge})^T]$, and since $\text{E}[\xi_t\xi_s] = 0$ for any non-equal t and s and $\text{E}[\xi_t e_1] = 0$ for any $t > 1$, all off-diagonal elements of $\text{Cov}[\mathbf{Ge}]$ are zero. For the diagonal elements, all but one of them are effectively the same as before, so are clearly σ_e^2. It only remains then to show that $\text{E}[(-g\sum_{j=0}^{m}\rho^j\xi_{s+m+1-j})^2] = \sigma_e^2$. We show this as follows:

$$\begin{aligned} \text{E}[(-g\sum_{j=0}^{m}\rho^j\xi_{s+m+1-j})^2] &= \text{E}[g^2\sum_{j=0}^{m}\rho^{2j}\xi_{s+m+1-j}^2] && (2.59)\\ &= \sigma_e^2 g^2\sum_{j=0}^{m}\rho^{2j} && (2.60)\\ &= \sigma_e^2(\sum_{r=0}^{m}\rho^{2r})^{-1}\sum_{j=0}^{m}\rho^{2j} && (2.61)\\ &= \sigma_e^2. && (2.62) \end{aligned}$$

So, since we have $\text{Cov}[\mathbf{G}\boldsymbol{\epsilon}] = \sigma_e^2\mathbf{I}$ again, it follows that for a single gap of missing observations of size m the matrix $\mathbf{G}$ defined in Equation 2.53 is the unique matrix which gives $\mathbf{V^{-1}} = \mathbf{G}^T\mathbf{G}$ for the GLS technique assuming an AR(1) covariance structure in the errors. Furthermore, if there are multiple gaps of missing observations of any size m, it is easy to

see how $\mathbf{G}$ is modified (i.e. the $-g\rho^{m+1}$ and g matrix elements are inserted in the proper locations in place of $-\rho$ and 1 respectively).

For those more concerned with theory, we make one final note. From this verification, it can be seen with a little extra thought how the matrix $\mathbf{G}$ could be found in an almost systematic way. This gives insight into how this matrix was discovered in the first place. In (Savin and White, 1978) this verification is not shown explicitly; the matrix is just given. So, this insight could have been lost.

2.4.3.2 Estimating ρ and σ_e^2

In order to estimate $\hat{\boldsymbol{\beta}}$ and $\text{Cov}[\hat{\boldsymbol{\beta}}]$ in the Prais-Winsten procedure it should be noted that estimates of ρ and σ_e^2 are needed (ρ is the only parameter $\mathbf{V^{-1}}$ depends upon, and σ_e^2 shows up right in the equation for $\text{Cov}[\hat{\boldsymbol{\beta}}]$ if you recall that $a = \sigma_e^2$). In practice, these are in fact estimated with an OLS regression on the AR(1) model we have written for the error terms. This model is stated again here for convenience as,

$$e_i = \rho e_{i-1} + \xi_i, \qquad \xi_i \sim [0, \sigma_e^2] \qquad i = 2, 3, ..., n. \tag{2.63}$$

Using the general notation presented before for OLS estimation, the variable ρ would be the only regression coefficient and is estimated with Equation 2.13, while σ_e^2 is estimated with Equation 2.15. To be completely specific, the matrix $\mathbf{X}$ (vector in this case of a single regressor) and data vector $\mathbf{y}$ for this MLR model are given as,

$$\mathbf{X} = \mathbf{X}_p = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{n-1} \end{bmatrix}, \qquad \mathbf{y} = \mathbf{y}_p = \begin{bmatrix} \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}, \tag{2.64}$$

where we have defined the variables $\mathbf{X}_p$ and $\mathbf{y}_p$ to represent these vectors for this specific regression, and $\epsilon_1, ..., \epsilon_n$ are the residuals for some MLR model estimation on the data. So, we note that in order to run this OLS estimate for ρ (and to estimate σ_e^2), residuals for an MLR model (likely using OLS estimation itself) must first be obtained. For notation, let us define this OLS estimate of ρ as ρ^*. So, we see that we have,

$$\rho^* = (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \mathbf{X}_p^T \mathbf{y}_p. \tag{2.65}$$

We have now given all the tools needed to use the Prais-Winsten FGLS method of iteratively calculating a GLS estimate until it is found that the matrix $\mathbf{V}$ varies by less than some tolerance. Again, the FGLS procedure first runs the OLS and uses its residuals to estimate ρ (with the OLS given in this section). This ρ^* gives an estimate of $\mathbf{V}$, so the GLS can then be run. The GLS is then run iteratively, using the residuals to get new estimates of ρ each time, until we decide to stop when V (or just ρ^* for this Prais-Winsten case) varies by less than some tolerance. This algorithm is given in a table in Section 2.4.6 after the next two sections on the model fit and confidence intervals.

2.4.4 The Model Fit

In this section, we give the model fit (the point estimates of the data $y_1, ..., y_n$ that a model provides) for the MLR model. It should come as no surprise that for both the OLS and GLS estimation techniques this estimate is $\mathbf{X}\boldsymbol{\beta}^*$. We denote this model fit as $\mathbf{Y}^*$. The model fit estimator, which we define as $\hat{\mathbf{Y}}$, is given as,

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \tag{2.66}$$

and its excepted value is the model fit $\mathbf{Y}^*$,

$$\mathrm{E}[\hat{\mathbf{Y}}] = \mathbf{X}\mathrm{E}[\hat{\boldsymbol{\beta}}] = \mathbf{X}\boldsymbol{\beta}^* = \mathbf{Y}^*. \tag{2.67}$$

Taking the covariance, the covariance matrix of this estimator is found to be,

$$\mathrm{Cov}[\hat{\mathbf{Y}}] = \mathbf{X}\mathrm{Cov}[\hat{\boldsymbol{\beta}}]\mathbf{X}^T. \tag{2.68}$$

So, for OLS estimation we have,

$$\mathrm{Cov}[\hat{\mathbf{Y}}] = \sigma^2 \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \tag{2.69}$$

and for GLS estimation we have,

$$\text{Cov}[\hat{\mathbf{Y}}] = a^2\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T. \tag{2.70}$$

2.4.5 Confidence Intervals

In this section, we show how to quantify how precisely the MLR regression coefficients and model fit are estimated. This is done by creating confidence intervals.

2.4.5.1 Confidence Intervals on the Regression Coefficients

Determining confidence intervals for the regression coefficient estimates requires an additional assumption for the MLR model. Until now, the type of distribution of the error terms has not been specified, only its mean and covariance given. If these error terms are assumed to be Gaussian distributed, then $\mathbf{Y}$ is also Gaussian distributed, and therefore $\hat{\boldsymbol{\beta}}$ is also Gaussian distributed since it is a linear function of $\mathbf{Y}$ (because linear operations on Gaussian distributed random vectors result in Gaussian distributed random vectors). In practice, this is typically a good assumption because of the prevalence of Gaussian noise in nature. Some resources on MLR make this assumption immediately, but it has not technically been necessary until now. With this assumption the MLR model equation we wrote for OLS estimation becomes

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \qquad \mathbf{e} \sim N[0, \sigma^2\mathbf{I}], \tag{2.71}$$

and the MLR model equation we wrote for GLS estimation becomes

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \qquad \mathbf{e} \sim N[0, a^2\mathbf{V}]. \tag{2.72}$$

The only difference in these from the model equations presented previously is the "N" to denote that the random vector is Gaussian distributed with the specified mean and covariance.

First we will consider confidence intervals for OLS estimation. Recall that $\text{Cov}[\hat{\boldsymbol{\beta}}]$ is given by,

$$\text{Cov}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}, \tag{2.73}$$

for the OLS estimation. If we denote $(\mathbf{X}^T\mathbf{X})^{-1}_{ii}$ as the ith diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$ then the quantity $\sigma\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}$ is the square root of the ith diagonal element of $\text{Cov}[\hat{\boldsymbol{\beta}}]$, and this quantity is the standard deviation of $\hat{\beta}_i$ (if we define $\hat{\beta}_i$ as the ith element of $\hat{\boldsymbol{\beta}}$). Now, given this, the random variable Z,

$$Z = \frac{\hat{\beta}_i - \beta_i^*}{\sigma\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}}, \tag{2.74}$$

is the standard normal (Gaussian) random variable (Appendix G gives an understanding of why this is a standard normal random variable). So, this random variable could be used to create confidence intervals. However, this is typically not possible because the value of σ is rarely known in practice. But, an estimate of σ can be obtained from Equation 2.15 by taking the square root of the estimate σ^{2^*}. Lets likewise denote this estimate as σ^* and define its estimator as S. With this, we can write the random variable T,

$$T = \frac{\hat{\beta}_i - \beta_i^*}{S\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}}, \tag{2.75}$$

which turns out to be t distributed with $n - (k + 1)$ degrees of freedom. So, in practice we use this random variable to obtain confidence intervals on the regression coefficients. The $100(1 - \alpha)\%$ confidence interval for a regression coefficient β_i is therefore defined by the bounds,

$$\begin{aligned} \text{lower bound:} \quad & \beta_i^* - t_{\alpha/2,n-(k+1)}\sigma^*\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}} \\ \text{upper bound:} \quad & \beta_i^* + t_{\alpha/2,n-(k+1)}\sigma^*\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}, \end{aligned} \tag{2.76}$$

where $t_{\alpha/2,n-(k+1)}$ is the critical value of a t distribution with a significance level of $\alpha/2$ and $n - (k + 1)$ degrees of freedom.

If we wish to know the confidence level of the sign of β_i^*, which is something we will do when we report the results for the MLR model on the SOO data record, then we find the value of α that makes the following equation hold:

$$t_{\alpha,n-(k+1)} = \frac{|\beta_i^*|}{\sigma^* \sqrt{(\mathbf{X}^T\mathbf{X})_{ii}^{-1}}}. \tag{2.77}$$

Then, the confidence level of the sign is $100(1-\alpha)\%$ given the found α. Lastly for the OLS confidence interval, let us just define the estimate of $\text{Cov}[\hat{\boldsymbol{\beta}}]$ as,

$$\text{Cov}[\hat{\boldsymbol{\beta}}]^* = \sigma^{2^*}(\mathbf{X}^T\mathbf{X})^{-1}, \tag{2.78}$$

since it is clearly of interest to calculate with the square of its diagonal elements showing up in our confidence interval.

For the GLS estimation in the Prais-Winsten procedure, the story is basically the same. We have,

$$\text{Cov}[\hat{\boldsymbol{\beta}}] = \sigma_e^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{2.79}$$

Again, we can not use the standard normal random variable for confidence intervals unless we know σ_e. Instead, we can use Equation 2.15 from the OLS regression described in Section 2.4.3.2 to estimate σ_e. Then we have the quantity,

$$\text{Cov}[\hat{\boldsymbol{\beta}}]^* = \sigma_e^{2^*}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}, \tag{2.80}$$

and the confidence intervals from the t distributed random variable become

$$\begin{aligned} \text{lower bound:} \quad & \beta_i^* - t_{\alpha/2,n-2}\sigma_e^* \sqrt{(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})_{ii}^{-1}} \\ \text{upper bound:} \quad & \beta_i^* + t_{\alpha/2,n-2}\sigma_e^* \sqrt{(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})_{ii}^{-1}}. \end{aligned} \tag{2.81}$$

The difference in the GLS Prais-Winsten case is that we have $n-2$ degrees of freedom instead of $n-(k+1)$. This works out this way because the OLS described in Section 2.4.3.2 is on data of length $n-1$ and there is only a single regression coefficient (i.e. $k=1$). Appendix H can be seen for more detail on these t distributions for creating confidence intervals in both the OLS and Prais-Winsten cases. Lastly, the $100(1-\alpha)\%$ confidence level for the sign of β_i^* in the Prais-Winsten case can be found from finding the α that makes the following equation hold:

$$t_{\alpha,n-2} = \frac{|\beta_i^*|}{\sigma_e^* \sqrt{(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})_{ii}^{-1}}}. \tag{2.82}$$

2.4.5.2 Confidence Intervals on the Model Fit

For confidence intervals on the model fit we have a similar situation to the regression coefficients. $\text{Cov}[\hat{\mathbf{Y}}]$ is not typically known in the OLS or GLS Prais-Winsten cases because σ^2 or σ_e^2 is not known. But similarly, with the estimates σ^{2^*} and $\sigma_e^{2^*}$ we can create t distributed random variables. We first define the following quantities:

$$\text{Cov}[\hat{\mathbf{Y}}]^* = \sigma^{2^*}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \tag{2.83}$$

for the OLS estimation and

$$\text{Cov}[\hat{\mathbf{Y}}]^* = \sigma_e^{2^*}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T, \tag{2.84}$$

for the GLS estimation and the Prais-Winsten estimation procedure. Then, for the OLS case for example, we can create a t distributed random variable with $n-(k+1)$ degrees of freedom as,

$$T = \frac{\hat{Y}_i - Y_i^*}{\sqrt{\text{Cov}[\hat{\mathbf{Y}}]_{ii}^*}}. \tag{2.85}$$

So, the $100(1-\alpha)\%$ confidence interval for a data point Y_i is defined by,

$$\begin{aligned} \text{lower bound:} \quad & Y_i^* - t_{\alpha/2,n-(k+1)}\sqrt{\text{Cov}[\hat{\mathbf{Y}}]_{ii}^*} \\ \text{upper bound:} \quad & Y_i^* + t_{\alpha/2,n-(k+1)}\sqrt{\text{Cov}[\hat{\mathbf{Y}}]_{ii}^*}. \end{aligned} \tag{2.86}$$

Similarly, the $100(1-\alpha)\%$ confidence interval for the Prais-Winsten estimation is defined by,

$$\begin{aligned} \text{lower bound:} \quad & Y_i^* - t_{\alpha/2,n-2}\sqrt{\text{Cov}[\hat{\mathbf{Y}}]_{ii}^*} \\ \text{upper bound:} \quad & Y_i^* + t_{\alpha/2,n-2}\sqrt{\text{Cov}[\hat{\mathbf{Y}}]_{ii}^*}. \end{aligned} \tag{2.87}$$

Where again, there is $n-2$ degrees of freedom for this one.

2.4.6 An Implementation of the Prais-Winsten Algorithm

In this section, an implementation of the Prais-Winsten MLR algorithm, the state-of-the-art for stratospheric ozone trend analysis, is given. This is done in the table below. The various variables in this table are defined and used in the previous sections, and the algorithm is the complete Prais-Winsten FGLS algorithm that has already been described. So, only a brief explanation of some of the implementation details are given here.

The tolerance variable t is the defined value that the difference between successive estimates of ρ has to be lower than before the procedure stops. When the while loop breaks based on this rule, the last GLS estimate (obtained using the last estimate of ρ) is stored in the $\boldsymbol{\beta}^*$ variable. This is essentially the primary result of the algorithm. This is then the appropriate time to calculate the quantities of $\text{Cov}[\hat{\boldsymbol{\beta}}]^*$, the model fit $\mathbf{Y}^*$, and $\text{Cov}[\hat{\mathbf{Y}}]^*$, since these are all likely desired by the modeller. So, we add these calculations to the algorithm in steps 14 through 17. As a byproduct, we also have estimates of the parameters ρ^* and $\sigma_e^{2^*}$ which are related to the autocorrelation in the data $\mathbf{y}$. Lastly, note that in steps 2 and 3 the variables α and ρ_{prev} are set so that there are a least two iterations of the while loop done. This means that there are at least two GLS estimates performed after the first OLS estimate.

Algorithm 1 Prais-Winsten MLR Algorithm

Input: Data $\mathbf{y}$, regressors $\mathbf{X}$, tolerance t for when the iteration should end

Output: GLS regression coefficient estimates $\boldsymbol{\beta}^*$, the covariance matrix $\text{Cov}[\hat{\boldsymbol{\beta}}]^*$, the model fit $\mathbf{Y}^*$, the covariance matrix $\text{Cov}[\hat{\mathbf{Y}}]^*$, estimates of the AR(1) model parameters ρ^* and $\sigma_e^{2^*}$

1: Perform the OLS estimate. Set: $\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

2: Set: $\alpha = 1$

3: Set: $\rho_{prev} = 2$

4: **while** $\alpha > t$ **do**

5: Calculate the residuals. Set: $\boldsymbol{\epsilon}^* = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*$

6: Construct $\mathbf{X}_p$ and $\mathbf{y}_p$ from the residuals $\boldsymbol{\epsilon}^*$

7: Perform the OLS estimate of ρ. Set: $\rho^* = (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{X}_p^T\mathbf{y}_p$

8: Construct the matrix $\mathbf{G}$ using $\rho = \rho^*$

9: Set: $\mathbf{V}^{-1} = \mathbf{G}^T\mathbf{G}$

10: Perform the GLS estimate. Set: $\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$

11: Set: $\alpha = |\rho_{prev} - \rho^*|$

12: Set: $\rho_{prev} = \rho^*$

13: **end while**

14: Calculate σ_e^2: $\sigma_e^{2^*} = \frac{(\mathbf{y}_p - \rho^*\mathbf{X}_p)^T(\mathbf{y}_p - \rho^*\mathbf{X}_p)}{n-2}$

15: Calculate the covariance matrix: $\text{Cov}[\hat{\boldsymbol{\beta}}]^* = \sigma_e^{2^*}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$

16: Calculate the model fit: $\mathbf{Y}^* = \mathbf{X}\boldsymbol{\beta}^*$

17: Calculate the covariance matrix: $\text{Cov}[\hat{\mathbf{Y}}]^* = \sigma_e^{2^*}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$

2.5 Prais-Winsten Estimation Results

In this section, the results from applying the Prais-Winsten MLR algorithm on the SOO relative anomaly data record are shown. For regressors, all of the indexes given in Section 2.2 are used. In the first section, the results for a single altitude-latitude time series are shown in detail and in the next section, we illustrate the results for all the altitude-latitude regions. The tolerance t (referring to Section 2.4.6) used for these is 0.01.

2.5.1 Example Time Series

In this section, the results at an altitude of 42.5 km and latitude region of 35° to 45° N are given. This altitude-latitude region has been selected because it is the region where a strong negative trend is seen from the beginning until about 1997 and a smaller positive looking trend afterward. This data is shown as the blue dots in Figure 2.9. The regression coefficient estimates $a_1, ..., a_8$ (referring to Equation 2.2) are,

$$\begin{bmatrix} -0.0062 & -0.0090 & 0.0066 & 0.0024 & 0.0017 & -0.0796 & 0.0293 & -0.0369 \end{bmatrix}, \tag{2.88}$$

and their standard deviations are (square root of the diagonal elements of $\text{Cov}[\hat{\boldsymbol{\beta}}]^*$),

$$\begin{bmatrix} 0.0026 & 0.0026 & 0.0029 & 0.0027 & 0.0032 & 0.0093 & 0.0049 & 0.0055 \end{bmatrix}. \tag{2.89}$$

The AR(1) coefficient ρ converged upon by the Prais-Winsten algorithm is 0.43, and the estimated variance of the AR(1) process $\sigma_e^{2^*}$ is 0.00094. The model fit for each index of the time series is shown as the orange line in Figure 2.9 with their 95% confidence intervals spanning the shaded region. We also show each of the 8 components of the model fit separately in Figure 2.10 (the sum is the model fit).

2.5.2 Results for the SAGE II/OSIRIS/OMPS Data Record

In this section, the same MLR results presented in the last section are shown, but for all altitude-latitude regions in the SOO data record. To do this, we illustrate the regression coefficients on heat maps with dimensions of altitude and latitude for each regressor. These are shown in Figures 2.11 and 2.12. Figure 2.11 is of the ozone influencing phenomena and Figure 2.12 is of the LINEAR PRE and LINEAR POST regressors multiplied by a number that converts them into the units of linear percent change per decade. This number is the change of each linear index term over 10 years multiplied by 100%. To give an idea of the certainty of the sign of these coefficients, we overlay solid, dashed, and dotted contours that show the percent confidence level that the estimated sign is correct. Recall that this can be

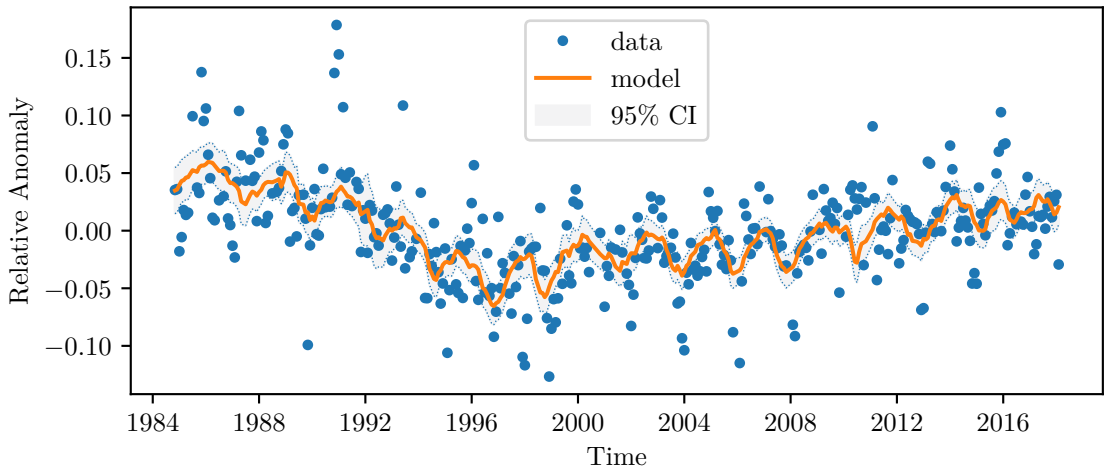


Figure 2.9: MLR model fit. SOO 42.5 km altitude 35° to 45° N latitude.

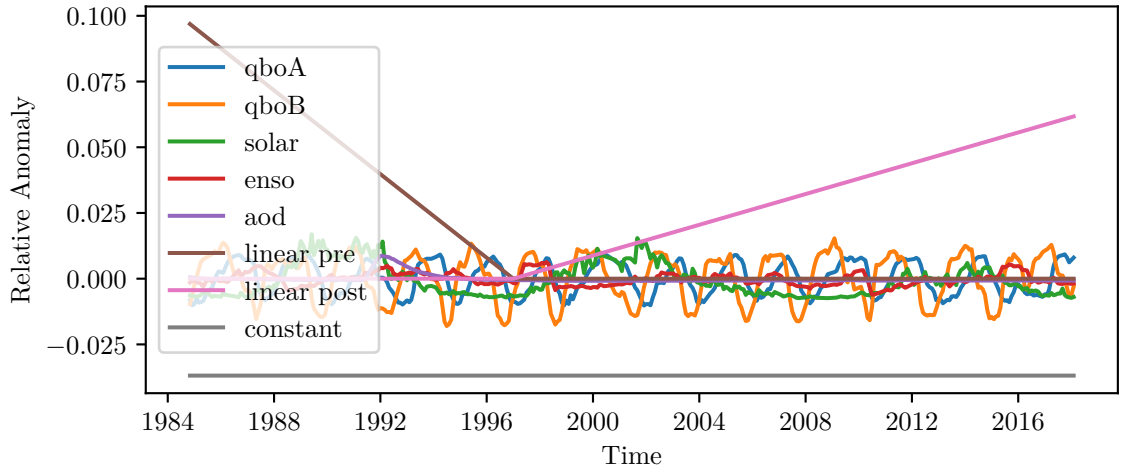


Figure 2.10: Components of MLR model fit. SOO 42.5 km altitude 35° to 45° N latitude.

calculated with Equation 2.82. The solid dashed and dotted lines represent the 80%, 90%, and 95% bounds respectively.

Lastly, in Figure 2.13 we show the converged AR(1) coefficients for each altitude-latitude region. Something that has not been done yet with this Prais-Winsten procedure, new to this thesis, is that we also show estimated 95% confidence intervals for the AR(1) coefficients in Figure 2.13. This is possible because if we assume the error term in the AR(1) model equation (Equation 2.32) is Gaussian distributed, then we can construct confidence intervals with t-critical values as we have already seen in this thesis. For instance, the variance of the OLS estimator of ρ can be estimated, as described with Equation 2.78, as,

$$Var(\hat{\rho})^* = \frac{\sigma_e^{2^*}}{\mathbf{X}_p^T \mathbf{X}_p}, \tag{2.90}$$

and the $100(1-\alpha)\%$ confidence interval is then given as,

$$\begin{aligned} \text{lower bound:} \quad & \rho^* - t_{\alpha/2,n}\sqrt{Var(\hat{\rho})^*} \\ \text{upper bound:} \quad & \rho^* + t_{\alpha/2,n}\sqrt{Var(\hat{\rho})^*}. \end{aligned} \tag{2.91}$$

In this chapter, the SOO data record has been described and the current state-of-the-art method of quantifying trends in stratospheric ozone has been presented. The following chapter begins to outline some background material for the developed DLM procedure of this thesis work.

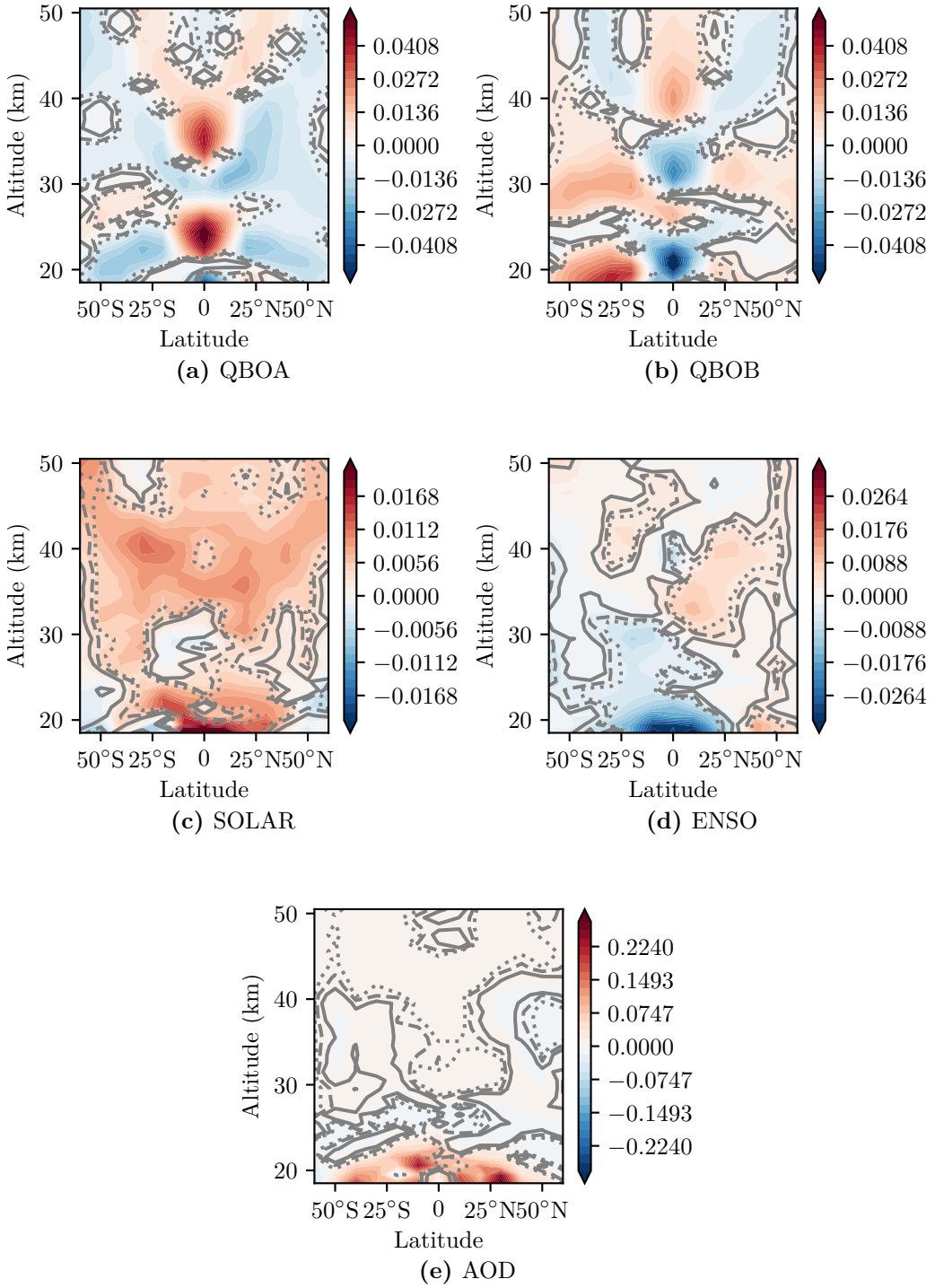
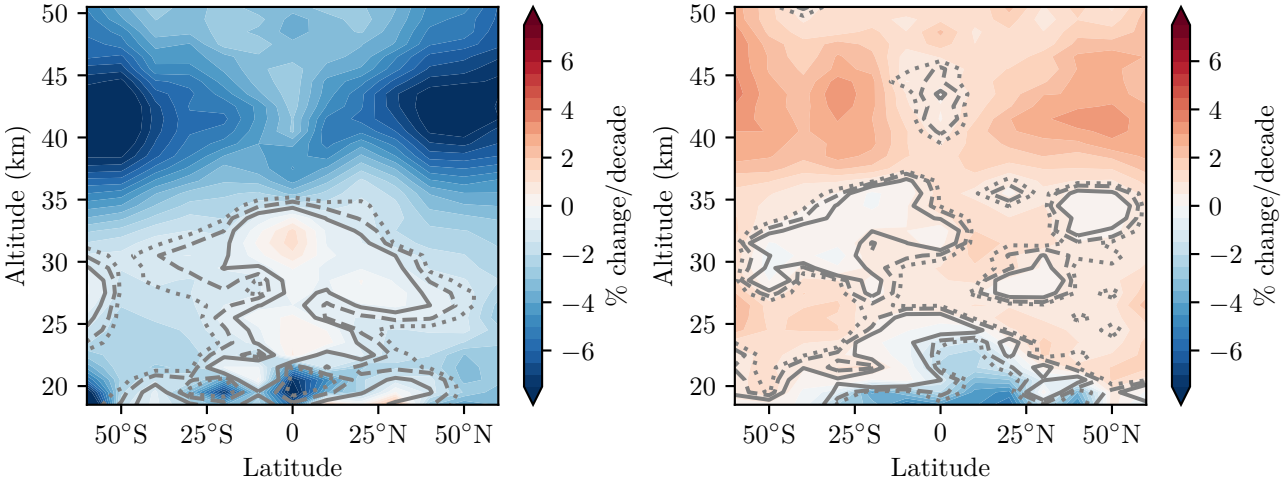


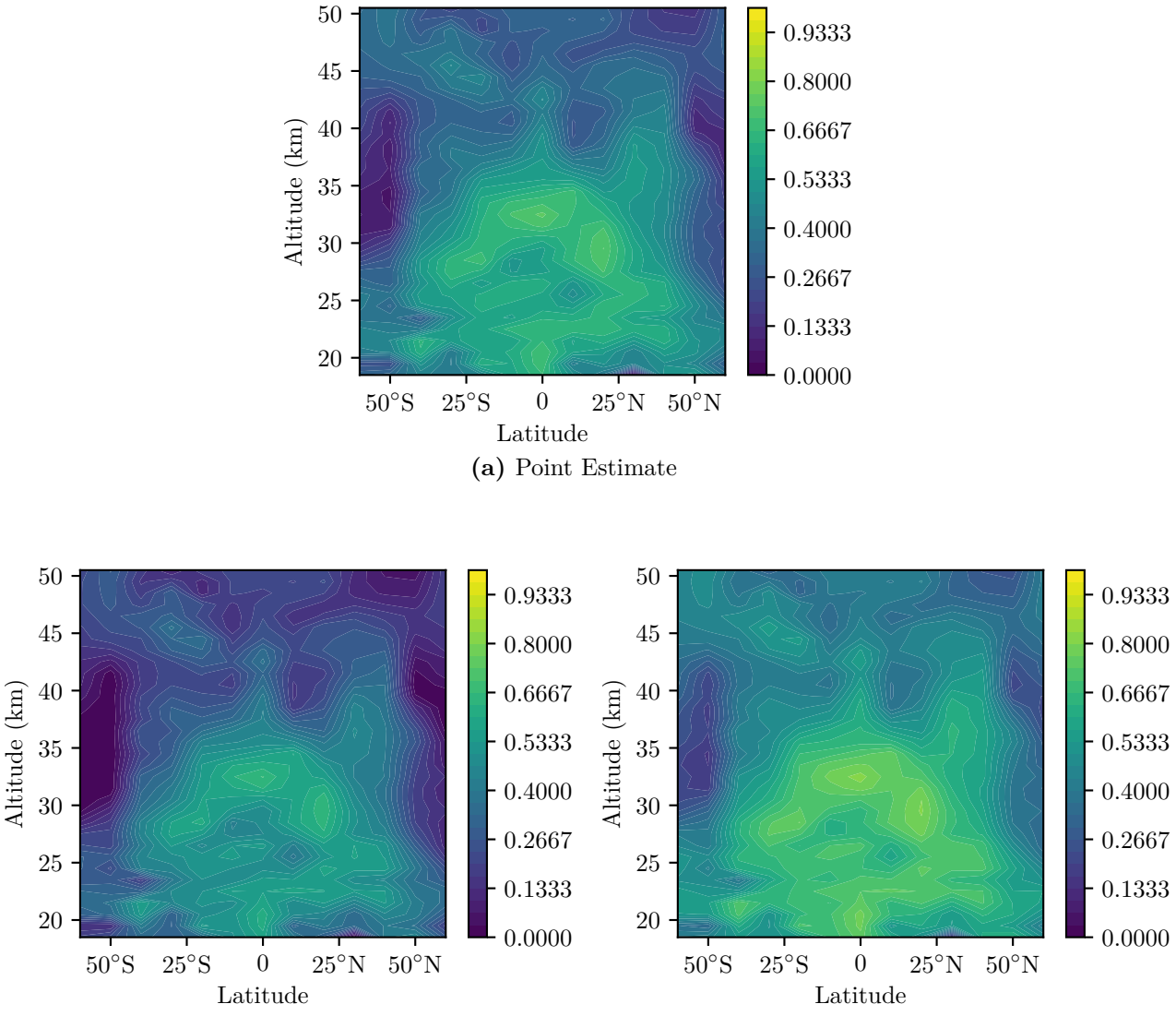
Figure 2.11: Ozone Influencing Phenomena Regression Coefficients

.



(a) LINEAR PRE **(b)** LINEAR POST

Figure 2.12: LINEAR PRE and LINEAR POST regression coefficients, converted to units of percent change per decade.



(a) Point Estimate

(b) 95% confidence interval lower bound

(c) 95% confidence interval upper bound

Figure 2.13: Estimated ρ from Prais-Winsten Procedure.

3 Markov Chain Monte Carlo

In this chapter, we cover the topic of Markov Chain Monte Carlo (MCMC). MCMC finds itself as an essential part of the DLM procedure developed in this thesis. It is used to estimate a DLM's unknown input parameters. Without this capability, we would be guessing on these parameters. This is all that will be said for motivation for now, and we treat the topic of MCMC very independently in this chapter, giving a good foundation into it, while not referring to any of the other topics of this thesis. It should be noted that MCMC does not come up again in this thesis until Chapter 5 when the DLM procedure is finally described. When reading this thesis completely through, this chapter and Chapter 4 can be read in either order.

At the core, MCMC methods are algorithms that allow for the drawing of a sample from any probability distribution. This is done by drawing individual random values, and repeating this until a sample of the desired size is obtained. Drawing random values from a probability distribution is not a straight forward thing. However, there exist methods to draw random values from Gaussian and uniform distributions. For any other probability distribution, MCMC methods can be used. As we will see in this chapter, all that is required to use them is actually the capability to draw random values from a uniform distribution, and usually, a Gaussian distribution, which as we said we already have the capability. In Section 3.1 we give the most fundamental of these algorithms called the Metropolis-Hastings algorithm. In Section 3.2 we give the MCMC concepts of convergence, visual diagnoses, burn-in, acceptance rate, and thinning. In Section 3.3 we discuss the parameter of the Metropolis-Hastings algorithm called the "proposal distribution", which is essential for practitioners to tune to their individual problem. In Section 3.4 we extend the Metropolis-Hastings algorithm to multivariate probability distributions. This extension is quite straight forward. Lastly, in Section 3.5 we give for reference a more complete Metropolis-Hastings algorithm

incorporating some of the topics covered throughout the chapter.

3.1 The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm, like all MCMC algorithms, allows for the sampling from a probability distribution. Consider a probability distribution function $p(x)$ that we wish to draw a sample from, and consider that we have a function $g(x)$ such that $p(x) \propto g(x)$. Also, consider any conditional probability distribution function that we define as $q(a|b)$. With these, the Metropolis-Hastings algorithm proceeds as shown in the table below. The result when complete is that the sequence $x_1, x_2, ..., x_m$ is a sample of size m from $p(x)$. However, it is important to note that this is only an approximation. What the theory behind the Metropolis-Hastings algorithm tells us (theory that we do not give in this thesis) is that we only have a limiting case result in that if m is infinite then the sequence $x_1, x_2, ..., x_m$ is a sample of infinite size from $p(x)$, without approximation.

Algorithm 2 Metropolis-Hastings

Input: Initial state, the number of iterations to perform m

Output: A sequence of m numbers $x_1, x_2, ..., x_m$ that is a sample (approximately) from $p(x)$

1: **Initialization:** Set: x_1 = initial state
2: **for** $i = 2 : m$ **do**
3: Set: x' = a random draw from the probability distribution $q(x|x_{i-1})$
4: Set: $\alpha = \frac{g(x')q(x_{i-1}|x')}{g(x_{i-1})q(x'|x_{i-1})}$
5: Set: u = a random number between 0 and 1
6: **if** $\alpha > u$ **then**
7: Set: $x_i = x'$
8: **else**
9: Set: $x_i = x_{i-1}$
10: **end if**
11: **end for**

For a brief discussion of this algorithm, we make several notes while also defining various

MCMC jargon terms. It can be seen that the "if" statement in the algorithm amounts to setting $x_i = x'$ if $\alpha \geq 1$ and otherwise, if $\alpha < 1$ setting $x_i = x'$ $100\alpha\%$ of the time and setting $x_i = x_{i-1}$ the other $100(1-\alpha)\%$ of the time. The value x' is commonly referred to as a "candidate" because it is a number that is sampled from $q(a|b)$ and then, in the jargon of MCMC, is either "accepted" or "rejected". It is accepted if the algorithm ends up setting $x_i = x'$ and rejected if not. The $q(a|b)$ distribution that randomly generates the candidates is referred to as the "proposal distribution" in light of this. You may also see it referred to less commonly as the "jumping distribution".

Now, other than the ability to perform arithmetic and compare numbers, take note that this algorithm requires two additional capabilities, that random draws from the proposal distribution can be taken and that random draws from a uniform distribution between 0 and 1 can be taken. Furthermore, note that the only required information about $p(x)$ to use this algorithm is that we are able to evaluate some function $g(x)$ that it is proportional to. Of course, if $p(x)$ can be evaluated itself it can be used as $g(x)$ in the algorithm.

There are a few more MCMC jargon terms that will be cleared up. One is that $p(x)$ is referred to as the "target distribution". We say this since it is the distribution we are "aiming" to draw a sample from. Another is that we refer to the sequence $x_1, x_2, ..., x_m$ as a "chain". Lastly, as the algorithm runs we say that the last value of x_i to be saved is the "state" of the MCMC "process" or MCMC "experiment". It should be noted also that this algorithm requires as input an initial state, which as we see is set to x_1, the first sample from the target distribution.

3.2 Convergence, Visual Diagnoses, Burn-in, Acceptance Rate, and Thinning

In this section, the MCMC topics of Convergence, Visual Diagnoses, Burn-in, acceptance rate, and thinning are defined and discussed.

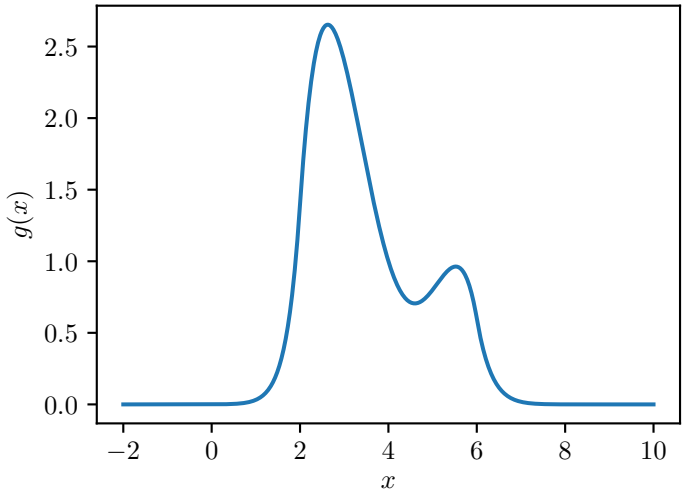


Figure 3.1: $g(x)$.

3.2.1 Convergence

Again, consider a defined function $g(x)$ which is proportional to the target distribution $p(x)$. Let us show a particular $g(x)$ in Figure 3.1 that is used as an example throughout the rest of this chapter. The MCMC chain is said to converge to $p(x)$ when the chain approaches equivalency with a true unbiased random sample from $p(x)$. As already stated prior, equivalency would require m to be infinity. So, in practice, we are only looking for when the chain is near equivalency or approaching equivalency. Broadly speaking, when this occurs a histogram generated by the values of the chain closely mimics the shape of the $g(x)$ or $p(x)$ curve. An example of this is shown in Figure 3.2 using the $g(x)$ function from Figure 3.1. The MCMC literature contains more sophisticated ways to assess convergence, but we give this simple way here for understanding and because it actually works fairly well in practice.

3.2.2 Visual Diagnoses

The first line of defense to assess the performance of an MCMC experiment is to look at visual diagnosis illustrations. We discuss two of such illustrations in this section. One of these we have already seen in Figure 3.2, which is just a histogram generated from the chain. Of course, normally in practice, we do not know the true shape of $p(x)$, so we cannot plot

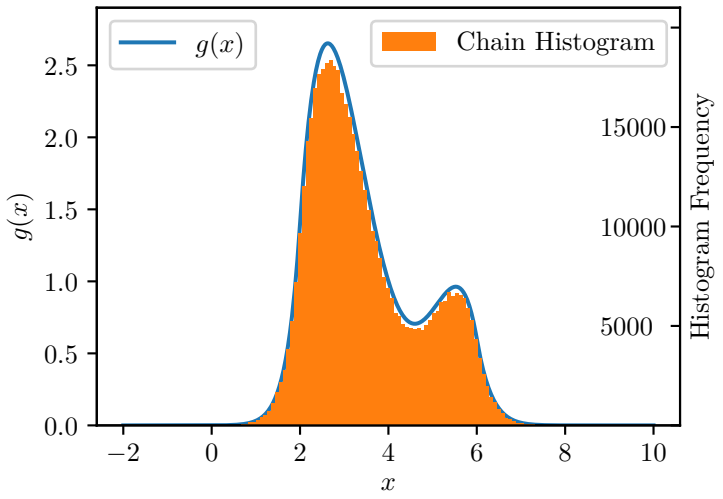


Figure 3.2: Histogram generated from the chain for a converged MCMC Experiment.

it with the histogram as shown in Figure 3.2. But, if we have just an idea of the shape of $p(x)$ this illustration can be useful in assessing the convergence of an MCMC experiment. For example, often we know that $p(x)$ is likely Gaussian or something close to Gaussian. So, we can simply look for the histogram to be something close to Gaussian, and if we see something which looks like a low amount of random draws from a Gaussian distribution, this may indicate that the chain is not converged yet and would become converged if ran for more iterations.

The second visual diagnosis is probably even more informative. It is called a trace plot and it is simply a line graph of the MCMC chain. We will show three example MCMC experiments with trace plots so that we can get an understanding of how to assess them.

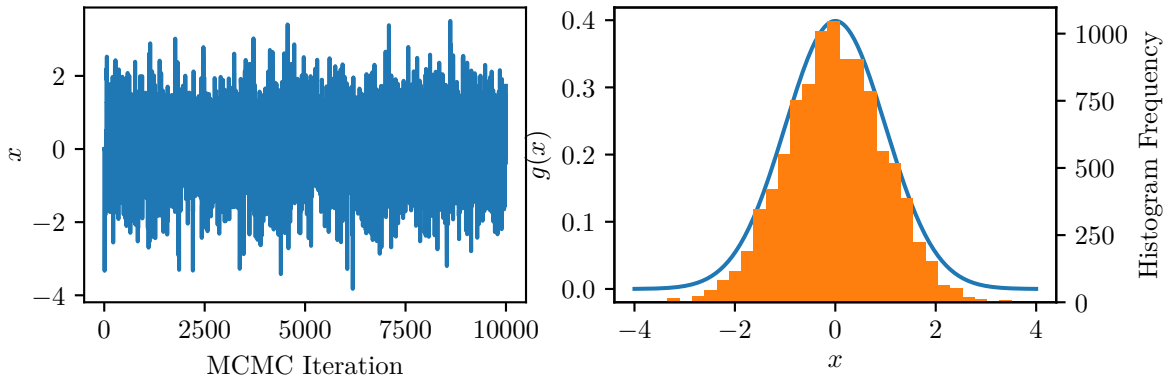
In Figure 3.3 the trace plots and histograms of three MCMC experiments are shown. The target distribution has been chosen to be Gaussian. We plot this Gaussian in blue with the orange histograms. In Experiment 1 we see a converged chain. We can see how the shape of the histogram approaches the shape of $p(x)$ and how the trace plot is solidly filled in when this happens. This is the type of trace plot we look for to indicate that we have a converged chain. In Experiment 2 we see from the histogram that the chain is not converged and from the trace plot that the value of the chain frequently remains the same for many consecutive iterations. For the Metropolis-Hastings algorithm, this happens, of course, when

the candidate x''s are a lot more often rejected than accepted. In Experiment 3 we have the opposite problem. We can see from the trace plot that the candidates are accepted at a high rate, but we see from the histogram that the chain is not converged. What is happening is that the candidates generated from the proposal distribution are all very close to the previous state, so it takes a large number of iterations before the chain can adequately explore all the probability regions of $p(x)$.

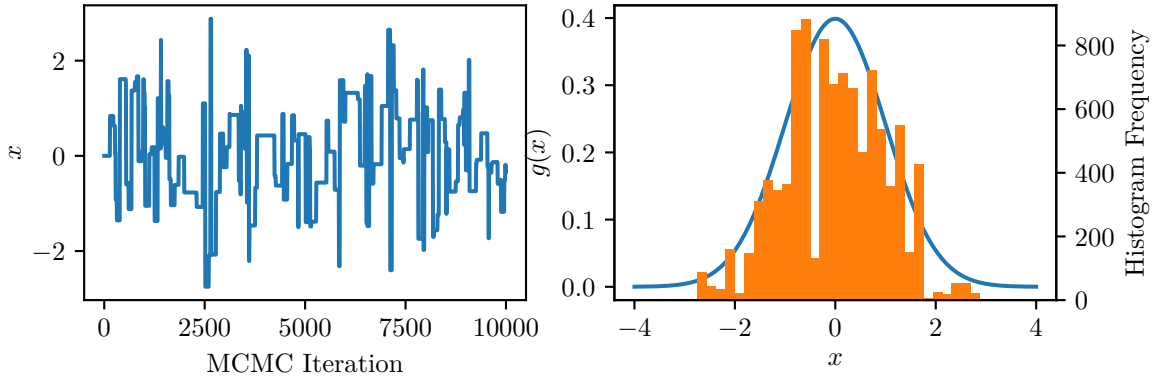
Experiments 2 and 3 show the two main problems encountered with MCMC in practice. For the Metropolis-Hastings algorithm, the parameter that controls how well the MCMC experiment avoids these two problems is the proposal distribution $q(a|b)$. So, when selecting the proposal distribution there is a balance to be made between these problems. In Section 3.3, we discuss this distribution more in-depth, give the most commonly used class of proposal distributions, and show how its parameter can be selected for the given target distribution of the MCMC experiment to lead to faster convergence. In fact, it is difficult to determine the optimal proposal distribution for a given target distribution. Instead, all practitioners typically aim for is something good enough so that the time (or number of iterations) it takes to obtain convergence is satisfactory. Selecting a better proposal distribution only means that the same result can be achieved with fewer iterations. To illustrate this, in Figure 3.4 we show the same MCMC experiment as Experiment 2 except that we run it for 30 times the amount of iterations. We can see that convergence is obtained, just as well as the convergence of Experiment 1 with 30 times fewer iterations.

3.2.3 Burn-in

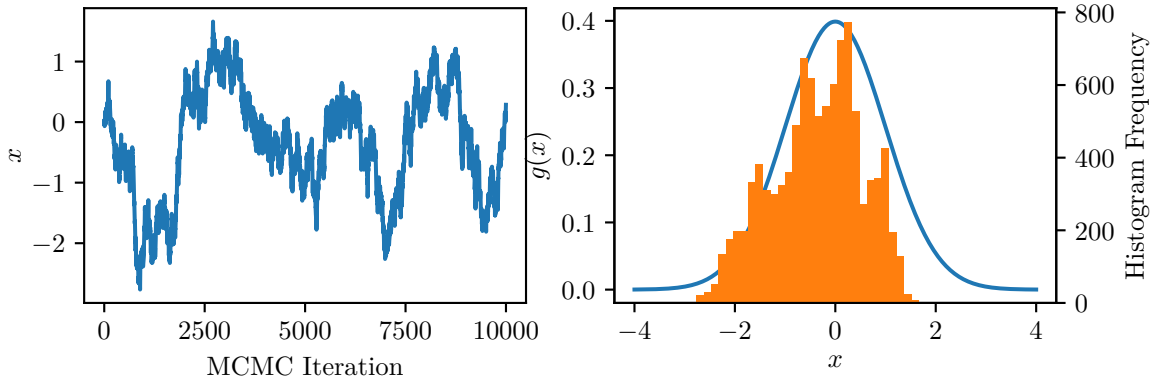
We will use trace plots for two additional MCMC experiments to illustrate an important concept for MCMC algorithms called a burn-in. As we have noted, the Metropolis-Hastings algorithm requires the specification of an initial state x_1. All MCMC algorithms share this characteristic actually (whether this is done under the hood or not). Figure 3.5 shows two MCMC experiments and illustrates what happens if the x_1 is selected in a high probability region of $p(x)$ versus what happens if it is selected in a very low probability region of $p(x)$. For these experiments we use the $g(x)$ function defined in Figure 3.1 and again show trace plots and histograms of the chain with $g(x)$ plotted as well. We see that when x_1 is selected



(a) Experiment 1



(b) Experiment 2



(c) Experiment 3

Figure 3.3: Three MCMC experiments. Left: trace plots. Right: histograms of the chains, blue: $g(x)$, orange: Chain Histograms.

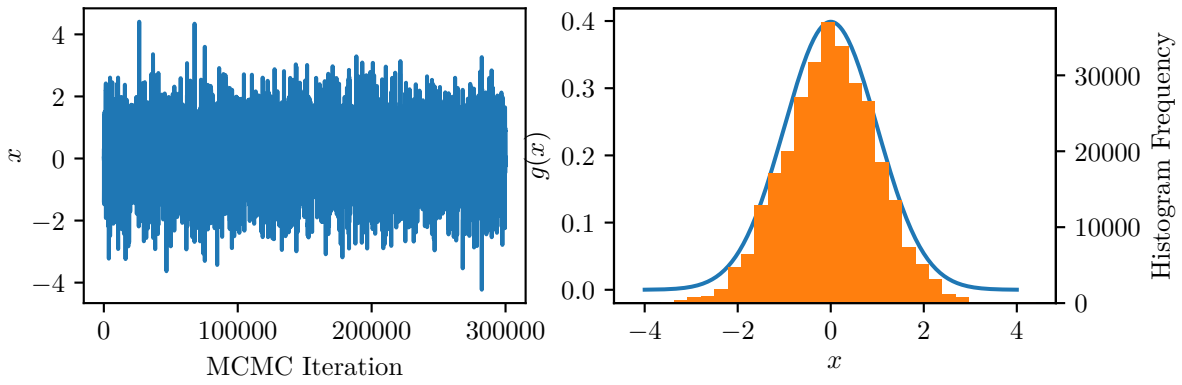


Figure 3.4: Experiment 2 in Figure 3.3 ran for 30 times more iterations.

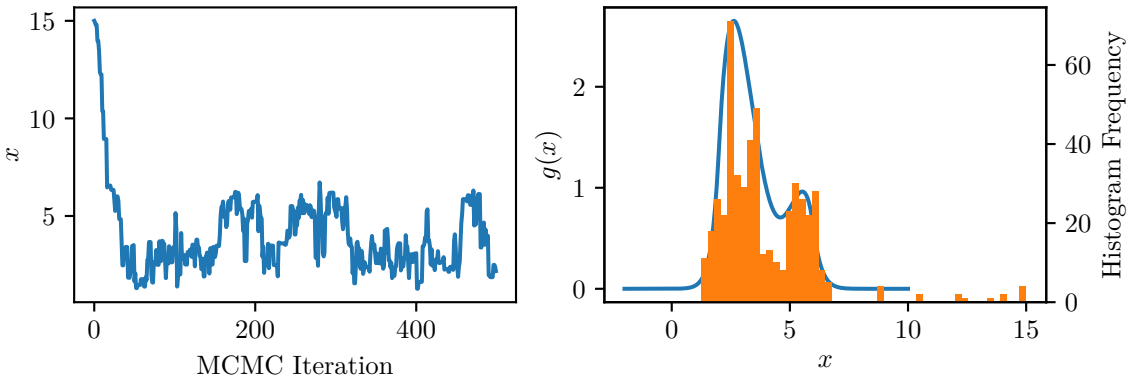
in a low probability region that the chain contains values at the beginning that are not easily encountered by the chain after it enters the high probability regions. Furthermore, we see that at the beginning the chain immediately begins to move towards the high probability regions. The same unwanted effect is not seen in the experiment for when x_1 is selected in a high probability region. The concept of a burn-in is to simply throw away some portion of the chain at the beginning, in hopes that the numbers for when the chain is still moving towards the high probability region are removed. In practice, the amount to throw away is typically chosen subjectively by the practitioner. To verify that their choice of this amount is sufficient for their MCMC experiment they can simply look at the trace plot.

3.2.4 Acceptance Rate

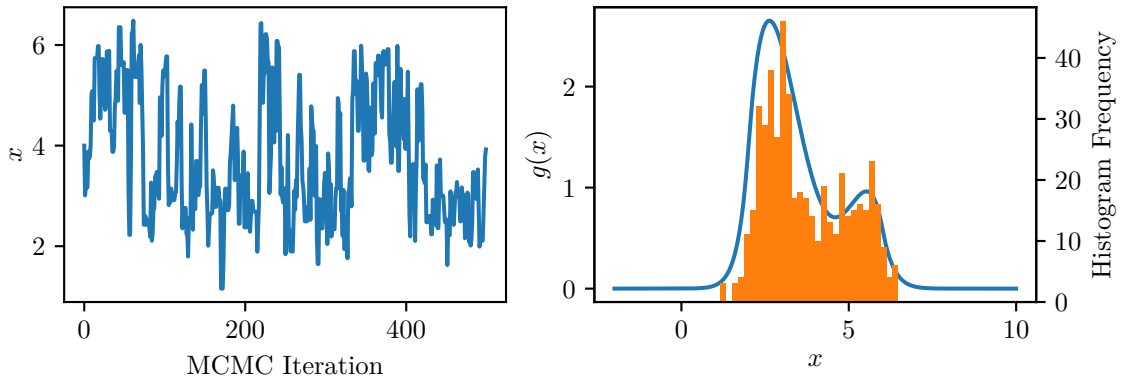
The acceptance rate for an MCMC experiment is defined as the fraction of the total $m-1$ candidate x''s that are accepted and become part of the chain.

3.2.5 Thinning

In this section, we discuss the concept of thinning an MCMC chain. Thinning is simply keeping only a fraction of the chain and throwing away the rest. Typically what MCMC practitioners do to thin a chain is to keep every nth element of it. But, there are of course other options. Instead of keeping every nth element we could throw away every nth element.



(a) x_1 in a low probability region



(b) x_1 in a high probability region

Figure 3.5: MCMC experiments explaining burn-in. Left: trace plots. Right: histograms of the chains, blue: $g(x)$, orange: Chain Histograms.

For example, throwing away every 3rd element could be reasonable for many MCMC setups so that two-thirds of the chain is kept. Another option could be something like keeping 2 in a row, then discarding 1, then keeping 1, then discarding 1, and then repeating this until the end of the chain so that three-fifths of the chain is kept. This is likely better than keeping three, then discarding 2, and repeating, for example.

The logic to modify the Metropolis-Hastings algorithm to include the capability for thinning is given in the final section of this chapter. Lastly, it needs to be noted that thinning a chain of size m is very rarely going to be better than keeping the non-thinned chain of size m. But, thinning a larger chain down so that you are left with a chain of size m is typically better than a non-thinned chain of the same size m. This means that essentially the only reason to thin a chain is to reduce the memory required to run the MCMC experiment on a computer.

3.3 The Proposal Distribution

In this section, the proposal distribution $q(a|b)$ used in the Metropolis-Hastings algorithm is discussed and potential choices for it are given. Recall that it is the distribution that is used to generate candidate x''s, and understand that it can be any distribution and the result of the Metropolis-Hastings algorithm is still valid. It is typically chosen subjectively by the practitioner. However, as we have already seen, not all proposal distributions will lead to acceptable results, and a well-selected proposal distribution can lead to much faster convergence. So, it is important for the practitioner to tune their choice of proposal distribution so that they get satisfactory results.

In fact, some of the more advanced MCMC methods aim to solve the problem of obtaining a strong proposal distribution by updating the proposal distribution as the MCMC experiment runs. These are called adaptive Metropolis algorithms. The tricky part about creating these algorithms is ensuring that they maintain the result that every MCMC algorithm should have, that for an infinite number of iterations the chain is equivalent to a sample from the target distribution. We do not look at these algorithms in this thesis. Instead, we only use the basic metropolis-hasting algorithm.

3.3.1 Gaussian Centered at the Current State

In this section, we present the most commonly used class of proposal distributions for the Metropolis-Hastings algorithm. It is a Gaussian distribution centered at b such that,

$$q(a|b) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(a-b)^2}{2\sigma^2}}. \tag{3.1}$$

The practitioner then chooses the standard deviation σ to be suitable for the specific problem. With this proposal distribution, we see that the candidates are selected from a Gaussian distribution centered at the current state of the MCMC experiment (since in the candidate proposal step of the Metropolis-Hastings algorithm we have $b = x_{i-1}$ and x_{i-1} equal to the current state). So, this means that the candidate is equally likely to be larger or smaller than the current state and that there is a higher probability that the candidate will be further away from the current state with a larger selected standard deviation σ. As a side note, for this particular proposal distribution we see that $q(a|b)/q(b|a) = 1$. Therefore, step 4 in the Metropolis-Hastings algorithm can be simplified to $\alpha = g(x')/g(x_{i-1})$.

As a second side note, it is from this class of proposal distributions that the proposal distribution gets its secondary name of a "jumping distribution". When the chain moves from one state to another this is sometimes referred to as a "jump" in the MCMC jargon. The center of this Gaussian proposal distribution jumps along with a change of state, so it is also called a jumping distribution.

Now, to illustrate how the choice of σ affects convergence we show three MCMC experiments with different choices of σ, using the function $g(x)$ shown in Figure 3.1. The standard deviations selected are 0.05, 1, and 50. The trace plots and histograms of each experiment are shown in Figure 3.6. For the $\sigma = 0.05$ experiment, we see from the trace plot that the chain wanders the probability space slowly, so much so that it is not even able to find the smaller local maximum of $p(x)$. This is similar to the situation of Experiment 2 in Figure 3.3. For the $\sigma = 50$ experiment, we see from the trace plot that regularly the chain goes for a large number of iterations where no candidate x' is accepted. These are seen as the horizontal line segments in the plot. This is similar to Experiment 3 in Figure 3.3. For the $\sigma = 1$ experiment, we see from the histogram and the trace plot that the chain has converged

fairly well to $p(x)$, much better than the other two experiments at least. So, it can be seen now that when using this class of proposal distributions, the practitioner tries to select the standard deviation so that they get the fastest convergence. This can be done through trial and error with shorted MCMC experiments.

3.3.2 Value Constrained

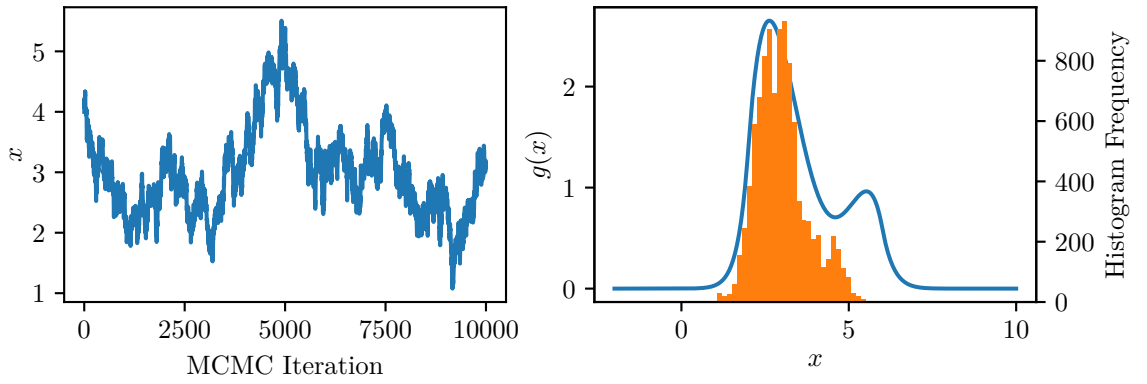
Another general option for proposal distributions is that they can be constrained to be between two values. The practitioner may choose to do this if they know for sure that x does not lay outside these values. For example, if we choose a uniform proposal distribution, we could have $q(a|b)$ as,

$$q(a|b) = \begin{cases} constant & d \geq a \geq c \\ 0 & a < c \text{ or } a > d \end{cases}, \tag{3.2}$$

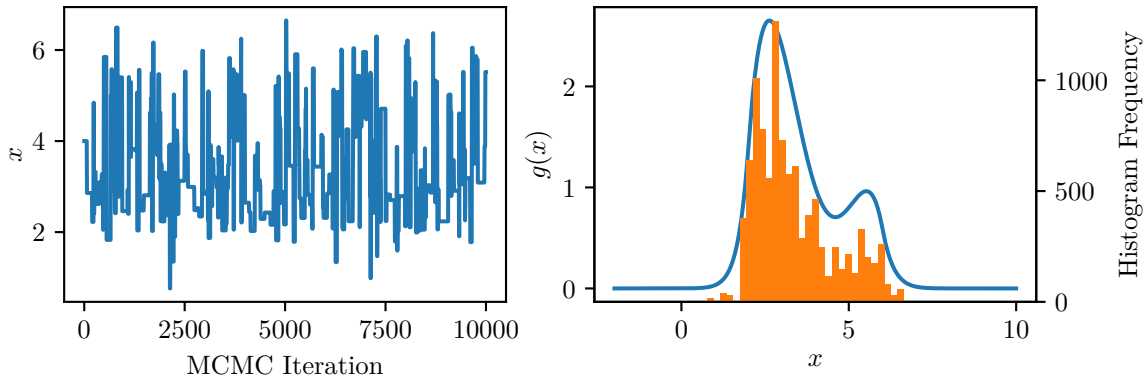
where the values c and d are the constraints. All values between c and d for this proposal distribution are equally likely to be proposed as candidates. In fact, a uniform distribution is clearly not practical to use when it is not value constrained. For another example, consider value constraining the Gaussian proposal distribution from the previous section to be between two values such that,

$$q(a|b) = \begin{cases} Ce^{-\frac{(a-b)^2}{2\sigma^2}} & d \geq a \geq c \\ 0 & a < c \text{ or } a > d \end{cases}. \tag{3.3}$$

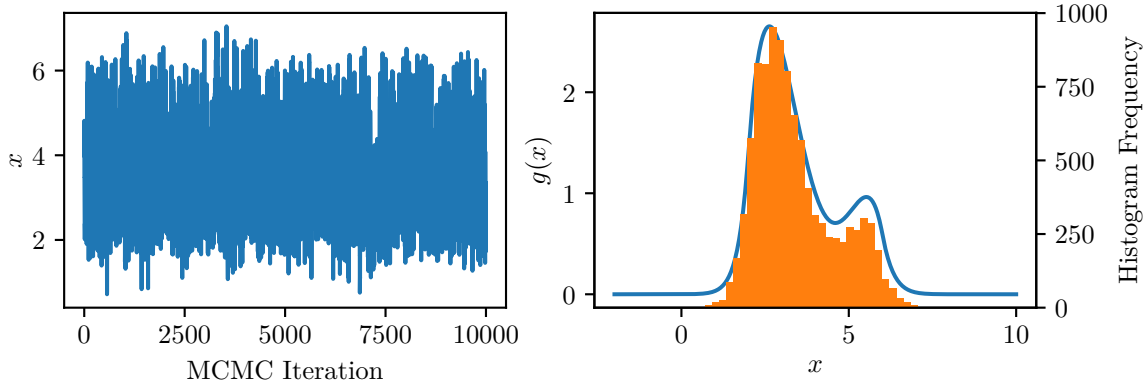
Such a function is called a truncated normal distribution. For a truncated normal distribution the normalizing constant C is calculable with a Gauss error function. It should be noted that for this truncated version we no longer have the simplification of $q(a|b)/q(b|a) = 1$. Lastly, we note that we can also consider value constrained proposal distributions to only have one bound in a sense if we choose either c as negative infinity or d as positive infinity.



(a) $\sigma = 0.05$



(b) $\sigma = 50$



(c) $\sigma = 1$

Figure 3.6: MCMC experiments with different choices of σ for the Gaussian centered at the current state proposal distribution. Left: trace plots. Right: histograms of the chains, blue: $g(x)$, orange: Chain Histograms.

3.3.3 Independence Sampler

Another class of proposal distributions is called independence samplers. These are characterized by $q(a|b) = q(a)$. In words, this means that the distribution that proposes candidates in the Metropolis-Hastings algorithm does not depend on the previous state. As an example, consider a Gaussian proposal distribution centered at some particular value, and that this value does not move throughout the entire MCMC experiment. When such a proposal distribution is used, it is noted in (Gilks et al., 1995) that this either works very well or very poorly depending on how well $q(a)$ covers the target distribution $p(x)$. This is intuitive since if $q(a)$ does not cover regions of $p(x)$ well then it will take long for the sequence to encounter these regions because $q(a)$ will not propose them as candidates often, and if it does cover $p(x)$ then the sequence is pretty much guaranteed to see all necessary regions of $p(x)$. If $q(a)$ covers a wide region much larger than $p(x)$ however, we start to get the situation of Experiment 3 in Figure 3.3 where too many candidates are rejected. So, this class of proposal distributions is basically only useful if the target distribution is known fairly well beforehand, in which case it can be very good. But, it can also be risky in that the chain can more easily miss regions of the target distribution without the practitioner noticing.

3.4 The Multivariate Metropolis-Hasting Algorithm

The extension of the Metropolis-Hastings algorithm from one variable to multiple is pretty simple and not a lot about the algorithm changes or needs to be discussed. We wish to draw a sample from a multivariate probability distribution $p(\mathbf{x})$ where we have a function $g(\mathbf{x})$ such that $p(\mathbf{x}) \propto g(\mathbf{x})$. We define the resulting sequence as $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m$ where each $\mathbf{x}_i$ is a vector of length n and the proposal distribution as $q(\mathbf{a}|\mathbf{b})$ where $\mathbf{a}$ and $\mathbf{b}$ are each length n. The Metropolis-Hastings algorithm is the same as the univariate case except for the x's are all vectors. The result at the end is that the sequence $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m$ is approximately a sample of size m from $p(\mathbf{x})$, with m as infinity giving no approximation. We show this slight modification of the algorithm in the table below.

Algorithm 3 Multivariate Metropolis-Hastings

Input: Initial state, the number of iterations to perform m

Output: A sequence of m numbers $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m$ that is a sample (approximately) from the probability distribution $p(\mathbf{x})$

1: **Initialization:** Set: $\mathbf{x}_1$ = initial state
2: **for** $i = 2 : m$ **do**
3: Set: $\mathbf{x}'$ = a random draw from the probability distribution $q(\mathbf{x}|\mathbf{x}_{i-1})$
4: Set: $\alpha = \frac{g(\mathbf{x}')q(\mathbf{x}_{i-1}|\mathbf{x}')}{g(\mathbf{x}_{i-1})q(\mathbf{x}'|\mathbf{x}_{i-1})}$
5: Set: u = a random number between 0 and 1
6: **if** $\alpha > u$ **then**
7: Set: $\mathbf{x}_i = \mathbf{x}'$
8: **else**
9: Set: $\mathbf{x}_i = \mathbf{x}_{i-1}$
10: **end if**
11: **end for**

3.4.1 The Proposal Distribution

We now discuss multivariate proposal distributions. A simplification that can be made is that if we choose the random variables of the proposal distribution to be independent such that,

$$q(\mathbf{a}|\mathbf{b}) = q(a_1|\mathbf{b})q(a_2|\mathbf{b})...q(a_n|\mathbf{b}) \tag{3.4}$$

$$= \prod_{j=1}^{n} q(a_j|\mathbf{b}), \tag{3.5}$$

where a_j is the jth element of $\mathbf{a}$ and then we relax the dependence of these distributions on $\mathbf{b}$ furthermore such that they are only dependent on the similar element of $\mathbf{b}$, we have,

$$q(\mathbf{a}|\mathbf{b}) = \prod_{j=1}^{n} q(a_j|b_j). \tag{3.6}$$

Then, the ratio $q(\mathbf{x}_{i-1}|\mathbf{x}')/q(\mathbf{x}'|\mathbf{x}_{i-1})$ seen in the Metropolis-Hastings algorithm becomes

$$\frac{q(\mathbf{x}_{i-1}|\mathbf{x}')}{q(\mathbf{x}'|\mathbf{x}_{i-1})} = \prod_{j=1}^{n} \frac{q((\mathbf{x}_{i-1})_j|(\mathbf{x}')_j)}{q((\mathbf{x}')_j|(\mathbf{x}_{i-1})_j)}, \tag{3.7}$$

where $(\mathbf{x})_j$ is the jth element of $\mathbf{x}$. This is a common choice, because with this we can now effectively specify proposal distributions separately for each variable in $\mathbf{x}$, using the single variable proposal distributions discussed in Section 3.3, where the discussion had still applies.

However, it should be noted that this is certainly not necessary, and it actually restricts us slightly in our choices for proposal distributions. For example, a multivariate Gaussian proposal distribution centered at $\mathbf{b}$ could be used with covariance between the variables. Such a proposal distribution cannot possibly be created using Equation 3.6. To be clear, if the practitioner chose there to be no covariance between variables then this multivariate Gaussian proposal distribution could be equivalently formulated using Equation 3.6 with multiple single variable Gaussian proposal distributions centered at b. But, with a multivariate Gaussian proposal distribution with covariance between its variables, this is a more complicated proposal distribution that can clearly not be made using equation 3.6.

3.4.2 Visual Diagnoses

For visual diagnoses in the multivariate case, we can similarly look at trace plots and histograms. This time we have them for each of the n components of $\mathbf{x}$. With these illustrations, we essentially look for the same things already discussed with univariate MCMC experiments. Examples of this were given throughout this chapter.

Let us make one additional note. Consider the n sequences from each of the n components of $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m$. The note is that these sequences are an approximate sample from the marginal distributions of $p(\mathbf{x})$. So, this means that the histograms generated by these sequences will closely resemble the marginal distributions when the chain converges.

3.5 Final Algorithm

In this section, we show for reference a final algorithm which is an implementation of the Multivariate Metropolis-Hastings algorithm with the capability of performing a specified

burn-in and specified thinning. This is shown in the table below. For a quick explanation, we require as input to the algorithm the number of burn-in iterations b. The burn-in part of the algorithm is steps 1 through 9. At the end of step 9, the state at the end of the burn-in is stored in the variable $\mathbf{x}_1$. The rest of the algorithm is identical to before except for the thinning logic. For the thinning, we require as input to the algorithm an array of Booleans of length m (to total number of iterations of the MCMC experiment) that specify which states to save to the chain and which to throw away. We denote this array as *thin*. For the thinning logic, we introduce the extra index variable j, which lags further behind the index variable i as states are elected to not be saved, and the variable $\mathbf{x}_{state}$, which keeps track of the current state of the MCMC process.

In this chapter, we have given an introduction to the topic of MCMC, which is a vital part of the DLM procedure developed in this thesis. In the next chapter, the DLM is introduced.

Algorithm 4 Multivariate Metropolis-Hastings with Burn-in and Thinning

Input: Initial x at the beginning of the burn-in, the number of iterations to perform m, the number of burn-in iterations to perform b, a Boolean array of length m indicating which sample points in the chain to keep *thin*.

Output: A sequence of s numbers $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_s$, where s is the number of True's in *thin*, that is a sample (approximately) from the probability distribution $p(\mathbf{x})$.

Initialization: Set: $\mathbf{x}_1$ = initial x at the beginning of the burn-in
for b iterations **do**
    Set: $\mathbf{x}'$ = a random draw from the probability distribution $q(\mathbf{x}|\mathbf{x}_1)$
    Set: $\alpha = \frac{g(\mathbf{x}')q(\mathbf{x}_1|\mathbf{x}')}{g(\mathbf{x}_1)q(\mathbf{x}'|\mathbf{x}_1)}$
    Set: u = a random number between 0 and 1
    if $\alpha > u$ **then**
        Set: $\mathbf{x}_1 = \mathbf{x}'$
    end if
end for
Set: $\mathbf{x}_{state} = \mathbf{x}_1$
Set: $j = 1$
if *thin*[1] **then**
    Set: $j = j + 1$
end if
for $i = 2 : m$ **do**
    Set: $\mathbf{x}'$ = a random draw from the probability distribution $q(\mathbf{x}|\mathbf{x}_{state})$
    Set: $\alpha = \frac{g(\mathbf{x}')q(\mathbf{x}_{state}|\mathbf{x}')}{g(\mathbf{x}_{state})q(\mathbf{x}'|\mathbf{x}_{state})}$
    Set: u = a random number between 0 and 1
    if *thin*[i] **then**
        if $\alpha > u$ **then**
            Set: $\mathbf{x}_j = \mathbf{x}'$
            Set: $\mathbf{x}_{state} = \mathbf{x}'$
        else
            Set: $\mathbf{x}_j = \mathbf{x}_{state}$
        end if
        Set: $j = j + 1$
    else if $\alpha > u$ **then**
        Set: $\mathbf{x}_{state} = \mathbf{x}'$
    end if
end for

4 Dynamic Linear Model

In this chapter, we introduce the dynamic linear model (DLM), the fundamental model of the procedure developed in this thesis for quantifying trends in time series. In Section 4.1 we introduce the recursive least squares (RLS) algorithm. This algorithm is simply a recursive version of the OLS estimation formula for the MLR model. It finds its way in as the first section of this chapter because of its similarity in operation to the DLM estimation equations and because we reference its theory further in this chapter. It can be thought of as a transition topic between MLR and DLMs. As mentioned in the introduction of this thesis, this section can be skipped over for a faster read. In Section 4.2 we introduce two unique DLMs that serve as an introduction to DLMs in general. In Section 4.3 we give a general form of the DLM that encompasses all DLMs that we would want to create. In Section 4.4 we give the recurrence relations that are used to estimate the "states" of the DLM. This is the analogous task of estimating the regression coefficients for the MLR model. In Section 4.5 we go into the theory behind the DLM estimation equations, giving several derivations of them using different statistical criteria. Now, for modelling a time series with a DLM, typically the approach is to have a set of fundamental DLMs that model different characteristics of the time series and then "superimpose" these DLMs to create a single DLM. The resulting single DLM is able to model all the characteristics together. So, in Section 4.6 we show a number of fundamental DLMs and then superimpose them to construct a DLM that is well suited for ozone trend analysis. Lastly, in Section 4.7 we describe what the ozone model looks like for the DLM we have constructed.

We note that the reader can safely skip Section 4.5, which is on the theoretical background of the recurrence relations that are used to estimate the states, and instead just use this section as a reference to the theory if desired at a later date. This is a rather large and detail-filled section. So, the reader can do what most people who use these recurrence relations in

practice do, and just accept that they result in statistically optimal estimates for DLMs and consider them as a tool for obtaining these statistically optimal estimates. An understanding of the theoretical underpinnings behind them is not necessary for the subsequent sections and chapters of this thesis.

4.1 Recursive Least Squares

The recursive least squares (RLS) algorithm for estimating $\boldsymbol{\beta}$ for the MLR model will now be shown. This algorithm yields the same answer as the OLS estimate given in Equation 2.13, but it provides an efficient way to update estimates of $\boldsymbol{\beta}$ when new data becomes available. For instance, consider a real-time application where data is sampled at a frequency of 1 Hz. Estimates of $\boldsymbol{\beta}$ can be updated efficiently every second with the RLS algorithm by just incorporating the new single piece of data, rather than performing the matrix multiplications with $\mathbf{X}$ over again. However, in the context of this thesis, the main reason why RLS is being shown is not for its usefulness in real-time applications, but for its comparison to the topic of DLMs. Namely, it has a similar recurrence relation algorithm that is used to estimate model parameters. The MLR estimation equations we have shown are sometimes referred to as "batch processing" in light of this algorithm which processes individual pieces of data.

4.1.1 Derivation

The RLS recurrence relation algorithm will now be derived from the OLS estimation formula, following closely to reference (Asada, 2006). Recall the OLS estimation formula,

$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{4.1}$$

and define $\mathbf{P} = (\mathbf{X}^T\mathbf{X})^{-1}$ and $\mathbf{B} = \mathbf{X}^T\mathbf{y}$ so that $\boldsymbol{\beta}^* = \mathbf{PB}$. Also, recall the definition of the matrix $\mathbf{X}$,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{k1} \\ x_{12} & \ddots & & \vdots \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ x_{1n} & \dots & \dots & x_{kn} \end{bmatrix}, \tag{4.2}$$

and define its rows as,

$$\boldsymbol{\varphi}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{bmatrix} \qquad i = 1, 2, ..., n. \tag{4.3}$$

It can be seen that,

$$\mathbf{P} = (\sum_{j=1}^{n} \boldsymbol{\varphi}_j \boldsymbol{\varphi}_j^T)^{-1}, \tag{4.4}$$

and,

$$\mathbf{B} = \sum_{j=1}^{n} y_j \boldsymbol{\varphi}_j. \tag{4.5}$$

Now, if $\mathbf{P}$ and $\mathbf{B}$ can be updated efficiently when new data becomes available then of course so can $\boldsymbol{\beta}^*$ by their multiplication. Let us define the ith update of $\mathbf{P}$, $\mathbf{B}$, and $\boldsymbol{\beta}^*$ to be $\mathbf{P}_i$, $\mathbf{B}_i$, and $\boldsymbol{\beta}_i^*$ respectively. Then, $\mathbf{B}_i$ may be written as,

$$\mathbf{B}_i = \sum_{j=1}^{i} y_j \boldsymbol{\varphi}_j \tag{4.6}$$

$$= \sum_{j=1}^{i-1} y_j \boldsymbol{\varphi}_j + y_i \boldsymbol{\varphi}_i \tag{4.7}$$

$$= \mathbf{B}_{i-1} + y_i \boldsymbol{\varphi}_i, \tag{4.8}$$

and $\mathbf{P}_i^{-1}$ may be written as,

$$\mathbf{P}_i^{-1} = \sum_{j=1}^{i} \boldsymbol{\varphi}_j \boldsymbol{\varphi}_j^T \tag{4.9}$$

$$= \sum_{j=1}^{i-1} \boldsymbol{\varphi}_j \boldsymbol{\varphi}_j^T + \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T \tag{4.10}$$

$$= \mathbf{P}_{i-1}^{-1} + \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T. \tag{4.11}$$

$\mathbf{P}_i$ can be calculated as follows by inverting the above equation and using the Woodbury matrix identity given in Appendix I:

$$\mathbf{P}_i = (\mathbf{P}_{i-1}^{-1} + \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T)^{-1} \tag{4.12}$$

$$= \mathbf{P}_{i-1} - \mathbf{P}_{i-1}\boldsymbol{\varphi}_i(1 + \boldsymbol{\varphi}_i^T \mathbf{P}_{i-1}\boldsymbol{\varphi}_i)^{-1}\boldsymbol{\varphi}_i^T \mathbf{P}_{i-1} \tag{4.13}$$

$$= \mathbf{P}_{i-1} - \frac{\mathbf{P}_{i-1}\boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T \mathbf{P}_{i-1}}{1 + \boldsymbol{\varphi}_i^T \mathbf{P}_{i-1}\boldsymbol{\varphi}_i}. \tag{4.14}$$

Notice that both $\mathbf{P}_i$ and $\mathbf{B}_i$ can be updated based on the previous estimates $\mathbf{P}_{i-1}$ and $\mathbf{B}_{i-1}$ with the above equations. Therefore, $\boldsymbol{\beta}_i^*$ can be updated with the equation $\boldsymbol{\beta}_i^* = \mathbf{P}_i \mathbf{B}_i$. This is fundamentally the RLS algorithm.

Additionally, this algorithm may be expressed in the following different form (this is actually the standard form that we will see used for the DLM recurrence relations, and is also more commonly used for the RLS algorithm):

$$\boldsymbol{\beta}_i^* = \boldsymbol{\beta}_{i-1}^* + \mathbf{K}_i(y_i - \boldsymbol{\varphi}_i^T \boldsymbol{\beta}_{i-1}^*). \tag{4.15}$$

This equation may be read as "the current estimation $\boldsymbol{\beta}_i^*$ is the sum of the previous estimation $\boldsymbol{\beta}_{i-1}^*$ and a linear operation $\mathbf{K}_i$ on the error of the previous estimation in predicting the current data point y_i". We see this because we interpret that $\boldsymbol{\varphi}_i^T \boldsymbol{\beta}_{i-1}^*$ is the best estimate of y_i given the previous estimation, or to think about it another way, given the data until index $i-1$. The linear operator $\mathbf{K}_i$ that makes this recurrence relation consistent with the RLS algorithm already shown is found to be (see Appendix J for these details),

$$\mathbf{K}_i = \frac{\mathbf{P}_{i-1}\boldsymbol{\varphi}_i}{1 + \boldsymbol{\varphi}_i^T \mathbf{P}_{i-1}\boldsymbol{\varphi}_i}. \tag{4.16}$$

This along with the update equation for $\mathbf{P}_i$ given in Equation 4.11 completes this other form of the RLS algorithm.

Recall that the original definition of $\mathbf{P}$ was $(\mathbf{X}^T\mathbf{X})^{-1}$ and that for the MLR model with OLS estimation the covariance of the regression coefficient estimator is $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. So, we just make the point here that the value $\mathbf{P}_i$ which shows up in the RLS recurrence relations is related to the error of the estimates.

Lastly, we note that by the way we have written the estimate as a function of the last estimate, it is natural that prior information can be incorporated with this procedure. The first data point is typical labelled y_1, and so $\boldsymbol{\beta}_0^*$ can be thought of as the best guess for the regression coefficients prior to analyzing any data and $\mathbf{P}_0$ as the confidence in this best guess. However, typically $\boldsymbol{\beta}_0^*$ is set to be a zero vector and $\mathbf{P}_0$ is set to be $\kappa\mathbf{I}$ where κ is some large number. This is meant to reflect that there is a great deal of uncertainty in the prior best guess of the regression coefficients. Doing this has the nice property that the resulting estimate of the regression coefficients at the end of the algorithm (i.e. $\boldsymbol{\beta}_n^*$) is the same as what you would find from the OLS batch processing estimation equation for $\boldsymbol{\beta}$.

4.1.2 Summary of Algorithms

The two different forms of the RLS algorithm derived in the above section are summarized in the tables below for easy reference.

Algorithm 5 RLS Algorithm 1

Input: $\mathbf{B}_0$: arbitrary, $\mathbf{P}_0$: arbitrary positive definite matrix, data $y_1, ..., y_n$

Output: The least squares estimate of regression coefficients incorporating data $y_1, ..., y_s$ (i.e. $\boldsymbol{\beta}_s^*$) for all $s = 1, ..., n$

1: **Initialization:** $\mathbf{B}_0 = \mathbf{B}_0$, $\mathbf{P}_0 = \mathbf{P}_0$

2: **for** $i = 1 : n$ **do**

3: $\quad \mathbf{B}_i = \mathbf{B}_{i-1} + y_i\boldsymbol{\varphi}_i$

4: $\quad \mathbf{P}_i = \mathbf{P}_{i-1} - \frac{\mathbf{P}_{i-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{P}_{i-1}}{1+\boldsymbol{\varphi}_i^T\mathbf{P}_{i-1}\boldsymbol{\varphi}_i}$

5: $\quad \boldsymbol{\beta}_i^* = \mathbf{P}_i\mathbf{B}_i$

6: **end for**

Algorithm 6 RLS Algorithm 2

Input: $\boldsymbol{\beta}_0^*$: arbitrary, $\mathbf{P}_0$: arbitrary positive definite matrix, data $y_1, ..., y_n$

Output: The least squares estimate of regression coefficients incorporating data $y_1, ..., y_s$ (i.e. $\boldsymbol{\beta}_s^*$) for all $s = 1, ..., n$

1: **Initialization:** $\boldsymbol{\beta}_0^* = \boldsymbol{\beta}_0^*$, $\mathbf{P}_0 = \mathbf{P}_0$

2: **for** $i = 1 : n$ **do**

3: $\quad \mathbf{K}_i = \frac{\mathbf{P}_{i-1}\boldsymbol{\varphi}_i}{1+\boldsymbol{\varphi}_i^T\mathbf{P}_{i-1}\boldsymbol{\varphi}_i}$

4: $\quad \boldsymbol{\beta}_i^* = \boldsymbol{\beta}_{i-1}^* + \mathbf{K}_i(y_t - \boldsymbol{\varphi}_t^T\boldsymbol{\beta}_{i-1}^*)$

5: $\quad \mathbf{P}_i = \mathbf{P}_{i-1} - \mathbf{K}_i\boldsymbol{\varphi}_i^T\mathbf{P}_{i-1}$

6: **end for**

4.1.3 Cost Function Theorem

Because it is informative for the DLM theory, we give the following theorem. For data $y_1, ..., y_n$ the RLS algorithm minimizes the following cost function $F_n(\boldsymbol{\beta})$:

$$F_n(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\varphi}_i^T\boldsymbol{\beta})^2 + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0^*)^T\mathbf{P}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0^*). \tag{4.17}$$

More specifically, what is meant is that the minima location of this function is at $\boldsymbol{\beta} = \boldsymbol{\beta}_n^*$. For the proof of this see Appendix K.

Recall that the cost function for OLS estimation is $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ (given by Equation 2.12). We see from linear algebra that

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{\varphi}_i^T\boldsymbol{\beta})^2. \tag{4.18}$$

So, the RLS cost function is identical to the OLS cost function minus an additional term that considers the prior information that the RLS uses. This is intuitive if we recall that OLS estimation does not use any prior information.

As a final note for the topic of RLS, we reference a simple modification to this algorithm called the weighted RLS algorithm. This modification weights certain data more than others in the estimation of the regression coefficients. So, it is related to the GLS version of MLR estimation, where what has been shown in this section is related to OLS. The weighted

RLS algorithm is given in Appendix L. We now transition to the topic of DLMs with the introduction of two unique DLMs.

4.2 Introductory Models

As an introduction to DLMs, two specific DLMs are given in this section. These are special cases of a more general DLM that is given in the next section. The two DLMs are called the multiple regression DLM, which should serve as a good introduction for the reason that it is similar to the MLR model that has already been shown in detail in this thesis, and a DLM referred to as the "first order polynomial DLM" or sometimes just the "local level DLM".

4.2.1 The Multiple Regression DLM

Recall the MLR model of Equation 2.6,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + e_i, \qquad e_i \sim [0, \sigma^2] \qquad i = 1, 2, ..., n, \tag{4.19}$$

and rewrite it more compactly as (using $\boldsymbol{\varphi_i}$ defined in Section 4.1),

$$Y_i = \boldsymbol{\varphi_i^T \beta} + e_i, \qquad e_i \sim [0, \sigma^2] \qquad i = 1, 2, ..., n. \tag{4.20}$$

The multiple regression DLM "model equation" is identical to this other than that it does not assume $\boldsymbol{\beta}$ to be constant for all index i. So, let us define $\boldsymbol{\beta_i}$ as the regression coefficient vector at index i. Furthermore, consider it to be a random vector that we choose to define as

$$\boldsymbol{\beta_i} = \boldsymbol{\beta_{i-1}} + \mathbf{w_i}, \qquad \mathbf{w_i} \sim N[0, \mathbf{W_i}], \tag{4.21}$$

where again, the notation $\mathbf{w_i} \sim N[0, \mathbf{W_i}]$ means $\mathbf{w_i}$ is a Gaussian distributed random vector with mean 0 and covariance matrix $\mathbf{W_i}$. Also, we will assume immediately that the e_i are Gaussian distributed, rather than waiting until we needed this assumption as we did in the MLR section of this thesis. So, the equations describing the multiple regression DLM are:

$$Y_i = \boldsymbol{\varphi}_\mathbf{i}^\mathbf{T}\boldsymbol{\beta}_\mathbf{i} + e_i, \qquad e_i \sim N[0, \sigma^2] \tag{4.22}$$

$$\boldsymbol{\beta}_\mathbf{i} = \boldsymbol{\beta}_\mathbf{i-1} + \mathbf{w_i}, \qquad \mathbf{w_i} \sim N[0, \mathbf{W_i}] \qquad i = 1, 2, ..., n. \tag{4.23}$$

The multiple regression DLM, like all DLMs, can be specified by two equations. What we call an observation equation, or model equation, is given by Equation 4.22 and what we call an evolution equation is given by Equation 4.23.

A statement about the statistical estimation procedure for the $\boldsymbol{\beta}_\mathbf{i}$ is deferred until the general DLM is defined. But take note that with this setup, we have constructed a model that is less constrained than the MLR model in that the regression coefficients can vary at each index, rather than requiring them to be the same. If one wanted to construct such a model for their data this is not something that is necessarily impractical either because as we have said there is an estimation procedure for DLMs that can be used to estimate these varying regression coefficients.

As another point, notice that if we choose $\mathbf{W_i} = 0$ then the evolution equation becomes uninteresting, telling us that the $\boldsymbol{\beta}_\mathbf{i}$ are all equal, and the model equation reduces to the MLR model equation. So, we consider the MLR model to be a special case of this multiple regression DLM. Furthermore, it is shown later in this thesis that for this special case the DLM estimation procedure effectively yields the same results as the MLR least squares estimation.

4.2.2 The Local Level DLM

The second introductory model we give is the local level DLM. The local level DLM is defined by the following observation and evolution equations:

$$Y_i = M_i + v_i, \qquad v_i \sim N[0, V_i] \tag{4.24}$$

$$M_i = M_{i-1} + w_i, \qquad w_i \sim N[0, W_i], \qquad i = 1, 2, ..., n. \tag{4.25}$$

By just looking at this definition, it is likely that the model's utility is not obvious. Suffice it to say that the random variable M_i represents the "level" or "background level" of the data at index i and that this background level can be estimated for each i with the DLM estimation

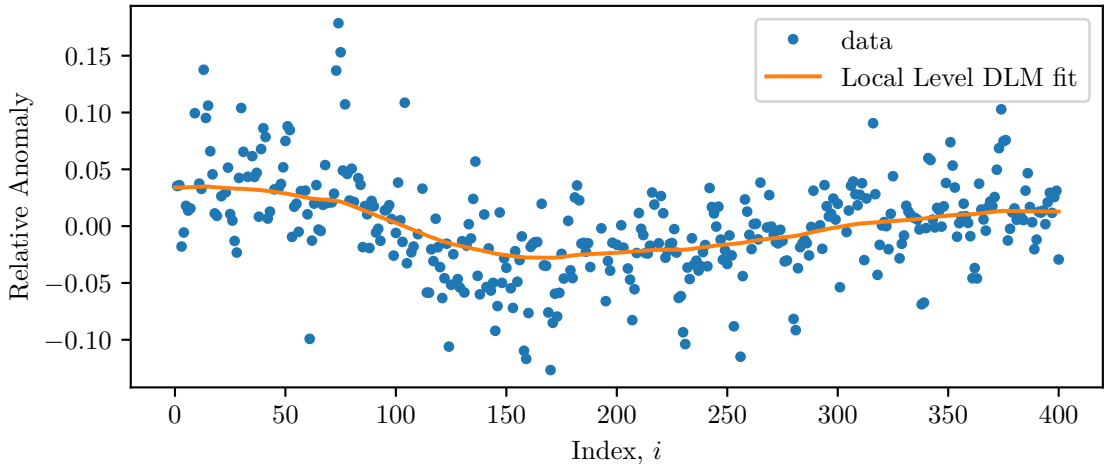


Figure 4.1: Local Level DLM Fit example. SOO 42.5 km altitude 35° to 45° N latitude.

procedure. To showcase this model, we give the estimated background level resulting from the DLM estimation procedure for some test data in Figure 4.1. This test data is actually the SOO MZM time series at 42.5 km altitude and 35°N to 45°N latitude that was seen before in this thesis. The variances V_i and W_i were defined to be 0.01 and 0.00001 respectively (these are required for the DLM estimation procedure).

4.3 A General Model

A more general DLM is presented in this section. Consider a DLM with the following observation and evolution equations:

$$Y_i = \mathbf{F}_i\mathbf{X}_i + v_i, \qquad v_i \sim N[0, V_i] \qquad i = 1, 2, ..., n, \tag{4.26}$$

$$\mathbf{X}_i = \mathbf{G}_i\mathbf{X}_{i-1} + \mathbf{w}_i, \qquad \mathbf{w}_i \sim N[0, \mathbf{W}_i] \qquad i = 1, 2, ..., n, \tag{4.27}$$

where Y_i is, again, the random variable that represents the data observation at index i, $\mathbf{X}_i$ is a random vector that is said to represent the "state of the system" at index i, and the vector $\mathbf{F}_i$, number V_i, and matrices $\mathbf{G_i}$ and $\mathbf{W}_i$ define the model relating Y_i to $\mathbf{X}_i$. We also refer

to values or estimates of the random vector $\mathbf{X}_i$ as the "state vector" or "parameter vector". Many authors choose not to distinguish between the random variable Y_i and observed values y_i, stating that it is clear from the context or not stating anything at all. We make sure to make this distinction in this thesis. Now, notice that this model encompasses the two DLMs presented previously. For the multiple regression DLM for example we just have $\mathbf{F}_i = \boldsymbol{\varphi}_i^T$ and $\mathbf{G_i} = \mathbf{I}$ and we consider $\mathbf{X}_i$ to be $\boldsymbol{\beta}_i$. We would like to solve the problem of estimating the state vector with this more general formulation since in doing so the problem is solved for all DLMs that can be described by it at once. It would be tedious to instead solve the problem for each DLM individually. We also note that we choose all the v_i's and $\mathbf{w}_i$'s (for all i) to be uncorrelated to each other.

To be slightly more general than the DLM above, consider the possibility of multiple measurements at each index i. For this, we define a random vector $\mathbf{Y}_i$ where each element represents one of the multiple measurements at index i. Using this in place of the single random variable Y_i, the observation and evolution equations become

$$\mathbf{Y}_i = \mathbf{F}_i\mathbf{X}_i + \mathbf{v}_i, \qquad \mathbf{v}_i \sim N[0, \mathbf{V}_i] \qquad i = 1, 2, ..., n, \tag{4.28}$$

$$\mathbf{X}_i = \mathbf{G}_i\mathbf{X}_{i-1} + \mathbf{w}_i, \qquad \mathbf{w}_i \sim N[0, \mathbf{W}_i], \qquad i = 1, 2, ..., n, \tag{4.29}$$

where $\mathbf{F}_i$ and $\mathbf{V}_i$ are now matrices as well, and again the $\mathbf{v}_i$'s and $\mathbf{w}_i$'s are uncorrelated. In time series applications this extra generalization is not typically useful since we usually just model one time series at a time. But, for engineering applications, where much of this theory arose, it is often necessary. We call the first set of equations in this section the univariate DLM and the above equations the general multivariate DLM or just the DLM. In this thesis we develop the DLM estimation procedure with the multivariate DLM generalization because this is how it is commonly done in the literature and because the univariate case is only a special case. This is done in the sections following this one.

Lastly, take notice that the DLM is specified by the four matrices $\mathbf{F}_i$, $\mathbf{G}_i$, $\mathbf{V}_i$, and $\mathbf{W}_i$ for all i. So, we will denote DLMs as "DLM $\{\mathbf{F}_i, \mathbf{G}_i, \mathbf{V}_i, \mathbf{W}_i\}$". Take for example the two introductory DLMs of Section 4.2. The multiple regression DLM is denoted as DLM $\{\boldsymbol{\varphi}_i^T, \mathbf{I}, V_i, \mathbf{W}_i\}$ where V_i and $\mathbf{W}_i$ could be any number and positive definite matrix for each

index i, and the local level DLM is denoted as DLM $\{1, 1, V_i, W_i\}$ where again, V_i and W_i could be any numbers for each index i. Most of the time $\mathbf{F}_i$ and $\mathbf{G}_i$ are really the defining matrices between classes of DLM models, and then $\mathbf{V}_i$ and $\mathbf{W}_i$ are selected in some manner for the specific application.

4.4 Estimation of the States

In this section, the recurrence relation algorithms for estimating the DLM state vectors are given. We first distinguish between three different circumstances for these estimations. Consider having observed the data $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$. Estimates of the state vector at index i where i is equal to n is called a filtered estimate, where i is less than n is called a smoothed (or interpolated) estimate, and where i is greater than n is called a prediction. This terminology stems from old engineering literature where noisy measurements are made of a signal and the "problems" of then inferring knowledge about the signal at times during the measurements, at the time of the last measurement, and at times after the last measurement are referred to as the smoothing problem, the filtering problem, and the prediction problem. To summarize, for data observed from index 1 to n, estimates of the state vector at index i are called:

$$\begin{aligned} i < n \quad & \text{a smoothed estimate,} \\ i = n \quad & \text{a filtered estimate, and} \\ i > n \quad & \text{a prediction.} \end{aligned}$$

Let us define $\tilde{\mathbf{y}}_n$ as the sequence $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ and similarly $\tilde{\mathbf{Y}}_n$ as the sequence $\mathbf{Y}_1, ..., \mathbf{Y}_n$. From the DLM estimation theory in the next section, we will see that we have,

$$p(\mathbf{x}_i|\tilde{\mathbf{y}}_n) \sim N[\mathbf{x}_i^n, \mathbf{P}_i^n]. \tag{4.30}$$

In words, this says that the probability distribution of the random vector $\mathbf{X}_i$ given the observed data $\tilde{\mathbf{y}}_n$ is Gaussian distributed with a mean that we have defined as $\mathbf{x}_i^n$ (=E$[\mathbf{X}_i|\tilde{\mathbf{y}}_n]$) and a covariance we have defined as $\mathbf{P}_i^n$ (=Cov$[\mathbf{X}_i|\tilde{\mathbf{y}}_n]$). Both $\mathbf{x}_i^n$ and $\mathbf{P}_i^n$ can be calculated in the cases of filtering, smoothing, and prediction from recurrence relations that we will

present shortly. So, with these being calculable we have a completely calculable probability distribution of the state vector at any index i. Since this is a Gaussian distribution, we, of course, consider $\mathbf{x}_i^n$ to be the point estimate of the state vector at any index i and $\mathbf{P}_i^n$ to describe the uncertainty in this estimate. The recurrence relations for calculating $\mathbf{x}_i^n$ and $\mathbf{P}_i^n$ for the cases of prediction, filtering, and smoothing are given in the following sections before we show the various statistical arguments for them in the subsequent sections.

4.4.1 Prediction

The recurrence relation for prediction is given by,

$$\mathbf{x}_i^n = \mathbf{G}_i \mathbf{x}_{i-1}^n, \tag{4.31}$$

where $i > n$. This equation can be recursively applied for each sequential prediction starting from the prediction one forward from the latest filtered estimate (i.e. $\mathbf{x}_{n+1}^n = \mathbf{G}_{n+1}\mathbf{x}_n^n$ where $\mathbf{x}_n^n$ is the filtering estimate made at index n). The corresponding covariance matrix recurrence relation is

$$\mathbf{P}_i^n = \mathbf{G}_i \mathbf{P}_{i-1}^n \mathbf{G}_i^T + \mathbf{W}_i, \tag{4.32}$$

and is applied similarly from the last filtering estimate covariance matrix $\mathbf{P}_n^n$.

4.4.2 Filtering

The recurrence relation for the filtered estimate is given by,

$$\mathbf{x}_i^i = \mathbf{G}_i \mathbf{x}_{i-1}^{i-1} + (\mathbf{G}_i \mathbf{P}_{i-1}^{i-1} \mathbf{G}_i^T + \mathbf{W}_i)\mathbf{F}_i^T (\mathbf{V}_i + \mathbf{F}_i (\mathbf{G}_i \mathbf{P}_{i-1}^{i-1} \mathbf{G}_i^T + \mathbf{W}_i)\mathbf{F}_i^T)^{-1} (\mathbf{y}_i - \mathbf{F}_i \mathbf{G}_i \mathbf{x}_{i-1}^{i-1}), \tag{4.33}$$

or written more compactly as,

$$\mathbf{x}_i^i = \mathbf{x}_i^{i-1} + \mathbf{P}_i^{i-1} \mathbf{F}_i^T (\mathbf{V}_i + \mathbf{F}_i \mathbf{P}_i^{i-1} \mathbf{F}_i^T)^{-1} (\mathbf{y}_i - \mathbf{F}_i \mathbf{x}_i^{i-1}), \tag{4.34}$$

using the prediction estimate $\mathbf{x}_i^{i-1}$ and covariance matrix $\mathbf{P}_i^{i-1}$ already defined. We note here that $\mathbf{x}_i^{i-1}$ along with its covariance matrix $\mathbf{P}_i^{i-1}$ is commonly referred to as the "one-step-

ahead" prediction. This equation can be recursively applied for each sequential estimation starting with the first data point $\mathbf{y}_1$. This starting point is given by,

$$\mathbf{x}_1^1 = \mathbf{x}_1^0 + \mathbf{P}_1^0\mathbf{F}_1^T(\mathbf{V}_1 + \mathbf{F}_1\mathbf{P}_1^0\mathbf{F}_1^T)^{-1}(\mathbf{y}_1 - \mathbf{F}_1\mathbf{x}_1^0), \tag{4.35}$$

where $\mathbf{x}_0^0$ and $\mathbf{P}_0^0$ (recall that $\mathbf{x}_1^0 = \mathbf{G}_1\mathbf{x}_0^0$ and $\mathbf{P}_1^0 = \mathbf{G}_1\mathbf{P}_0^0\mathbf{G}_1^T + \mathbf{W}_1$) are subjective prior estimates specified by the modeller. Typically these are chosen to be $\mathbf{x}_0^0 = 0$ and $\mathbf{P}_0^0 = \kappa\mathbf{I}$ where κ is a large number. This leads to the priors having little impact on the final estimations for all $\mathbf{x}_i^i$, especially when i is large. In other words it is a "non informative prior". From this estimate of $\mathbf{x}_1^1$ and the data $\mathbf{y}_2$, the next filtered estimate $\mathbf{x}_2^2$ can be found, and so on. The corresponding covariance matrix recurrence relation is

$$\mathbf{P}_i^i = \mathbf{P}_i^{i-1} - \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{V}_i + \mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T)^{-1}\mathbf{F}_i\mathbf{P}_i^{i-1}. \tag{4.36}$$

It should be stated that these recurrence relations for $\mathbf{x}_i^i$ and $\mathbf{P}_i^i$ are known as the Kalman Filter. This is a highly celebrated algorithm first published in engineering literature in 1960 that has found its use in many different applications. Let us define what is known as the Kalman gain matrix which shows up in the equations for both $\mathbf{x}_i^i$ and $\mathbf{P}_i^i$ as

$$\mathbf{K}_i = \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{V}_i + \mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T)^{-1}. \tag{4.37}$$

4.4.3 Smoothing

The recurrence relation for the smoothed estimate is given by,

$$\mathbf{x}_i^n = \mathbf{x}_i^i + \mathbf{P}_i^i\mathbf{G}_{i+1}^T{\mathbf{P}_{i+1}^i}^{-1}(\mathbf{x}_{i+1}^n - \mathbf{x}_{i+1}^i), \tag{4.38}$$

where $i < n$. Notice that $\mathbf{x}_{i+1}^i$ is a one-step-ahead prediction estimate and $\mathbf{x}_{i+1}^n$ is itself a smoothed estimate. This equation can be recursively applied starting with the smoothed estimate one index backwards from the last filtered estimate. That is, if we have a filtered estimate $\mathbf{x}_n^n$, first we can calculate $\mathbf{x}_{n-1}^n$ with

$$\mathbf{x}_{n-1}^n = \mathbf{x}_{n-1}^{n-1} + \mathbf{P}_{n-1}^{n-1}\mathbf{G}_n^T\mathbf{P}_n^{n-1}(\mathbf{x}_n^n - \mathbf{x}_n^n). \tag{4.39}$$

Then, we use this result to calculate $\mathbf{x}_{n-2}^n$, and so on. The corresponding covariance matrix recurrence relation is

$$\mathbf{P}_i^n = \mathbf{P}_i^i + \mathbf{P}_i^i \mathbf{G}_{i+1}^T {\mathbf{P}_{i+1}^i}^{-1} (\mathbf{P}_{i+1}^n - \mathbf{P}_{i+1}^i) {\mathbf{P}_{i+1}^i}^{-1} \mathbf{G}_{i+1} \mathbf{P}_i^i, \tag{4.40}$$

and is applied in the same way. For convenience, let us define an auxiliary parameter $\mathbf{J}_i$ that shows up in the equations for both $\mathbf{x}_i^n$ and $\mathbf{P}_i^n$ as,

$$\mathbf{J}_i = \mathbf{P}_i^i \mathbf{G}_{i+1}^T {\mathbf{P}_{i+1}^i}^{-1}. \tag{4.41}$$

4.4.4 Summary of DLM Estimation

The algorithms discussed in the previous three sections, which are simply the application of the given recurrence relations, for obtaining filtered, prediction, and smoothed estimates are summarized in the tables below for easy reference. We note that the smoothing and prediction algorithms require as inputs the filtered estimates of $\mathbf{x}_n^n$ and $\mathbf{P}_n^n$. As discussed above and shown in the tables below, these are the starting points for applying the smoothing and prediction recurrence relations. So, this means that to obtain any desired smoothed estimate or prediction the filtering algorithm must first be executed.

Algorithm 7 Filtering Algorithm

Input: $\mathbf{x}_0^0$: arbitrary, $\mathbf{P}_0^0$: arbitrary positive definite matrix, data $\mathbf{y}_1, ..., \mathbf{y}_n$

Output: The DLM filtered estimate $\mathbf{x}_i^i$ and covariance matrix $\mathbf{P}_i^i$ for all $i = 1, .., n$

1: **Initialization:** $\mathbf{x}_0^0 = \mathbf{x}_0^0$, $\mathbf{P}_0^0 = \mathbf{P}_0^0$

2: **for** $i = 1 : n$ **do**

3: $\quad \mathbf{x}_i^{i-1} = \mathbf{G}_i \mathbf{x}_{i-1}^{i-1}$

4: $\quad \mathbf{P}_i^{i-1} = \mathbf{G}_i \mathbf{P}_{i-1}^{i-1} \mathbf{G}_i^T + \mathbf{W}_i$

5: $\quad \mathbf{K}_i = \mathbf{P}_i^{i-1} \mathbf{F}_i^T (\mathbf{V}_i + \mathbf{F}_i \mathbf{P}_i^{i-1} \mathbf{F}_i^T)^{-1}$

6: $\quad \mathbf{x}_i^i = \mathbf{x}_i^{i-1} + \mathbf{K}_i (\mathbf{y}_i - \mathbf{F}_i \mathbf{x}_i^{i-1})$

7: $\quad \mathbf{P}_i^i = \mathbf{P}_i^{i-1} - \mathbf{K}_i \mathbf{F}_i \mathbf{P}_i^{i-1}$

8: **end for**

Algorithm 8 Prediction Algorithm

Input: filtered estimation and covariance matrix $\mathbf{x}_n^n$ and $\mathbf{P}_n^n$

Output: The DLM prediction $\mathbf{x}_i^n$ and covariance matrix $\mathbf{P}_i^n$ for any $i > n$

1: **Initialization:** $\mathbf{x}_n^n = \mathbf{x}_n^n$, $\mathbf{P}_n^n = \mathbf{P}_n^n$

2: **for** $i = n : \infty$ **do**

3: $\quad \mathbf{x}_i^n = \mathbf{G}_i \mathbf{x}_{i-1}^n$

4: $\quad \mathbf{P}_i^n = \mathbf{G}_i \mathbf{P}_{i-1}^n \mathbf{G}_i^T + \mathbf{W}_i$

5: **end for**

Algorithm 9 Smoothing Algorithm

Input: filtered estimation and covariance matrix $\mathbf{x}_n^n$ and $\mathbf{P}_n^n$

Output: The DLM smoothed estimate $\mathbf{x}_i^n$ and covariance matrix $\mathbf{P}_i^n$ for all $i = 1, ..., n-1$

1: **Initialization:** $\mathbf{x}_n^n = \mathbf{x}_n^n$, $\mathbf{P}_n^n = \mathbf{P}_n^n$

2: **for** $i = n - 1 : -1 : 1$ **do**

3: $\quad \mathbf{J}_i = \mathbf{P}_i^i \mathbf{G}_{i+1}^T {\mathbf{P}_{i+1}^i}^{-1}$

4: $\quad \mathbf{x}_i^n = \mathbf{x}_i^i + \mathbf{J}_i(\mathbf{x}_{i+1}^n - \mathbf{x}_{i+1}^i)$

5: $\quad \mathbf{P}_i^n = \mathbf{P}_i^i + \mathbf{J}_i(\mathbf{P}_{i+1}^n - \mathbf{P}_{i+1}^i)\mathbf{J}_i^T$

6: **end for**

4.4.5 DLM Estimation with Data Gaps

Handling gaps in the observed data with a DLM is pretty simple. We just notice that if there is missing data at some index s (i.e. $\mathbf{y}_s = NULL$), the filtering estimate $\mathbf{x}_s^s$ is equal to $\mathbf{x}_s^{s-1}$ by definition. If there is a larger data gap of size p, from index $s - p + 1$ to index s lets say, then the filtering estimate $\mathbf{x}_s^s$ is equal to $\mathbf{x}_s^{s-p}$. The same goes for $\mathbf{P}_s^s$ in relation to $\mathbf{P}_s^{s-p}$ by definition. So, the filtering algorithm given in the above section can easily be altered to handle gaps in the data as shown in the table below. If $\mathbf{y}_i$ exists then we apply the normal filtering estimate recurrence relation, and if it does not we just propagate the one-step-ahead predictions. The smoothing and prediction algorithms remain the same.

Algorithm 10 Filtering Algorithm

Input: $\mathbf{x}_0^0$: arbitrary, $\mathbf{P}_0^0$: arbitrary positive definite matrix, data $\mathbf{y}_1, ..., \mathbf{y}_n$

Output: The DLM filtered estimate $\mathbf{x}_i^i$ and covariance matrix $\mathbf{P}_i^i$ for all $i = 1, .., n$

Initialization: $\mathbf{x}_0^0 = \mathbf{x}_0^0$, $\mathbf{P}_0^0 = \mathbf{P}_0^0$
for $i = 1 : n$ **do**
  $\mathbf{x}_i^{i-1} = \mathbf{G}_i \mathbf{x}_{i-1}^{i-1}$
  $\mathbf{P}_i^{i-1} = \mathbf{G}_i \mathbf{P}_{i-1}^{i-1} \mathbf{G}_i^T + \mathbf{W}_i$
  if $\mathbf{y}_i$ exists **then**
    $\mathbf{K}_i = \mathbf{P}_i^{i-1} \mathbf{F}_i^T (\mathbf{V}_i + \mathbf{F}_i \mathbf{P}_i^{i-1} \mathbf{F}_i^T)^{-1}$
    $\mathbf{x}_i^i = \mathbf{x}_i^{i-1} + \mathbf{K}_i (\mathbf{y}_i - \mathbf{F}_i \mathbf{x}_i^{i-1})$
    $\mathbf{P}_i^i = \mathbf{P}_i^{i-1} - \mathbf{K}_i \mathbf{F}_i \mathbf{P}_i^{i-1}$
  else
    $\mathbf{x}_i^i = \mathbf{x}_i^{i-1}$
    $\mathbf{P}_i^i = \mathbf{P}_i^{i-1}$
  end if
end for

4.5 Estimation of the States Theory

In this section, we give theoretical background behind the prediction, filtering, and smoothing recurrence relations for the DLM. There are several statistical arguments/criteria that lead to the same recurrence relations. The prediction is rather straight forward, while filtering and smoothing are more complicated. In this section we show the following:

1. A simple proof for the prediction recurrence relations (Section 4.5.1);
2. The filtering and smoothing recurrence relations obtained from a procedure known as Recursive Bayesian Estimation (Section 4.5.2);
3. The filtering recurrence relations obtained as the minimum mean squared error (MMSE) estimate (Section 4.5.3);

4. The filtering and smoothing recurrence relations obtained as maxmimum likelihood estimates (MLEs) (Section 4.5.6); and

5. A cost function theorem, that is essentially a least squares cost function, that results in the prediction, filtering, and smoothing recurrence relations (Section 4.5.5).

References for these derivations are supplied within.

This work started in 1960 with the seminal paper of (Kalman, 1960) where the filtering recurrence relations were found with the criteria of MMSE. The filtering recurrence relations are known better as the Kalman filter for this reason. The original application of this filter was quite a bit different than what we are using it for in this thesis as a tool for time series analysis. It was developed for engineering applications where the problem was, very generally, to update an estimate for something in real-time based on noisy measurements that are related to the thing being estimated. As an example, the Kalman filter found itself as an essential part of NASA's Apollo Guidance Computer at this time. Only later was it realized that DLMs for time series using these recurrence relation tools can be useful. Many subsequent works followed from Kalman's seminal paper. Noteworthy among them is (Rauch et al., 1965) which shows that the criteria of MLE also results in the same filtering recurrence relations found by Kalman. With this, (Rauch et al., 1965) was able to extend the work of Kalman to find usable recurrence relations for the smoothing problem. These are the smoothing recurrence relations we have seen in this thesis and are known as the Rauch-Tung-Striebel smoother. So, we note that the Kalman filter and the Rauch-Tung-Striebel smoother are the two recurrence relations used in the DLM procedure (we do not really use the prediction recurrence relations in the DLM procedure unless you consider how they show up within the Kalman filter for the one-step-ahead predictions as using them). In this section, we assume some notions of probability and statistics are known to the reader, but we also go into details with a fair amount of depth.

4.5.1 Prediction Proof

First, we get the proof for the prediction recurrence relations out of the way since essentially it only involves the evolution equation. The quantities we want to calculate are $\mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{y}}_u]$ and

$\text{Cov}[\mathbf{X}_i|\tilde{\mathbf{y}}_u]$ for $i > n$.

For $\text{E}[\mathbf{X}_i|\tilde{\mathbf{y}}_u]$, we see that we have,

$$\begin{aligned}
\mathbf{x}_i^n &= \text{E}[\mathbf{X}_i|\tilde{\mathbf{y}}_n] && (4.42)\\
&= \text{E}[\mathbf{G}_i\mathbf{X}_{i-1} + \mathbf{w}_i|\tilde{\mathbf{y}}_n] && (4.43)\\
&= \mathbf{G}_i\text{E}[\mathbf{X}_{i-1}|\tilde{\mathbf{y}}_n] && (4.44)\\
&= \mathbf{G}_i\mathbf{x}_{i-1}^n, && (4.45)
\end{aligned}$$

and we note that this is only valid for $i > n$ because for $i \leq n$ it is no longer the case that

$$\text{E}[\mathbf{w}_i|\tilde{\mathbf{y}}_n] = 0. \quad (4.46)$$

This can be seen from the expanded observation equation presented in Equation M.7 in Appendix M. Because, in the case where $i > n$ $\mathbf{w}_i$ and $\mathbf{y}_n$ are independent, but for $i \leq n$ they are not. So, when they are independent we have $\text{E}[\mathbf{w}_i|\tilde{\mathbf{y}}_n] = \text{E}[\mathbf{w}_i] = 0$.

For $\text{Cov}[\mathbf{X}_i|\tilde{\mathbf{y}}_u]$, using the property of Appendix D (The property of Appendix D will no longer be referenced from this point on, the reader is assumed to know it now) we have,

$$\begin{aligned}
\mathbf{P}_i^n &= \text{Cov}[\mathbf{X}_i|\tilde{\mathbf{y}}_n] && (4.47)\\
&= \text{Cov}[\mathbf{G}_i\mathbf{X}_{i-1} + \mathbf{w}_i|\tilde{\mathbf{y}}_n] && (4.48)\\
&= \mathbf{G}_i\text{Cov}[\mathbf{X}_{i-1}|\tilde{\mathbf{y}}_n]\mathbf{G}_i^T + \mathbf{W}_i && (4.49)\\
&= \mathbf{G}_i\mathbf{P}_{i-1}^n\mathbf{G}_i^T + \mathbf{W}_i. && (4.50)
\end{aligned}$$

Similarly, this is only valid for $i > n$ where we have $\text{Cov}[\mathbf{w}_i|\tilde{\mathbf{y}}_n] = \text{Cov}[\mathbf{w}_i] = \mathbf{W}_i$. This completes the proof as these are the same recurrence relations for the prediction problem given in Section 4.4.1.

4.5.2 Recursive Bayesian Estimation

In this section, the filtering and smoothing recurrence relations for the general DLM are found by recursively applying Bayes theorem. For this, we cite (Petris et al., 2009) and

(Rodgers, 2000). (Rodgers, 2000) contains the derivation for filtering, while (Petris et al., 2009) contains the derivations for both filtering and smoothing. The approach used is known as Recursive Bayesian Estimation. First, we give an introduction to Bayes theorem.

4.5.2.1 Bayes Theorem

Let the probability of any event X be defined as $P(X)$. The traditional Bayes theorem is an equation for the probability of an event A occurring given that an event B has occurred (i.e. the conditional probability $P(A|B)$) in the form of:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4.51}$$

This can be proven by writing the joint probability of both A and B occurring as,

$$P(A, B) = P(B|A)P(A), \tag{4.52}$$

and equivalently as,

$$P(A, B) = P(A|B)P(B), \tag{4.53}$$

and then equating the two and dividing by $P(B)$.

There is a similar notion for probability distributions, which is also called Bayes theorem. The joint probability distribution $p(x, y)$ for the random variables X and Y is given as,

$$p(x, y) = p(x|y)p(y), \tag{4.54}$$

and equivalently as,

$$p(x, y) = p(y|x)p(x). \tag{4.55}$$

Then, equating the two and dividing by $p(y)$ yields Bayes theorem:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}. \tag{4.56}$$

However, this time we must be careful in the interpretation of this equation. Typically people think of $p(x|y)$ as a probability distribution of the random variable X given a known value y

of the random variable Y. However, we can also consider this function to have an unspecified value y and hence be a function of both variables, or even to have an unspecified value of y and a specified value of x so it is just a function of y. For Equation 4.56, suppose we think of $p(x|y)$ in the common way, as a function of x with a known value y, then the right-hand side of this equation must also be a function of only x. So, the $p(y|x)$ is actually a function of x for a fixed y in this scenario. This is called a likelihood function and we will write it as $\mathcal{L}(y|x)$ to indicate this. The functions values are a measure of how likely it is to observe the given observation y. For example, if we find that $\mathcal{L}(10|1) > \mathcal{L}(10|2)$, then for the given observation of $y = 10$ it is more likely that $x = 1$ than it is that $x = 2$. We also note that $p(x)$ is a function of x and $p(y)$ is only a constant in this scenario. Let us rewrite Bayes theorem with our new notation as,

$$p(x|y) = \frac{\mathcal{L}(y|x)p(x)}{p(y)}. \tag{4.57}$$

In the language of Bayesian inference, $p(x|y)$ is referred to as the posterior distribution, $\mathcal{L}(y|x)$ as the likelihood function, and $p(x)$ as the prior distribution. Since $p(y)$ is just a constant, it is often convenient to just write Bayes theorem as,

$$p(x|y) \propto \mathcal{L}(y|x)p(x). \tag{4.58}$$

We note that we do not require likelihood functions to integrate to 1 like we do for a probability distribution function.

It should be noted that the concepts of the likelihood function and the posterior distribution may seem the same based on the example of "$\mathcal{L}(10|1) > \mathcal{L}(10|2)$". It could be thought that the same could be said about $p(1|10) > p(2|10)$. The difference is that the posterior takes into account prior information (contained in $p(x)$) and the likelihood does not. Lastly, for random vectors $\mathbf{X}$ and $\mathbf{Y}$, the story is the same and we have Bayes theorem written as,

$$p(\mathbf{x}|\mathbf{y}) \propto \mathcal{L}(\mathbf{y}|\mathbf{x})p(\mathbf{x}). \tag{4.59}$$

4.5.2.2 A Note on Likelihood Functions

Before we find the DLM filtering recurrence relations with Recursive Bayesian Estimation, we make the following short discussion about likelihood functions. A likelihood function $\mathcal{L}(y|x)$ can still be written mathematically as if it were some probability distribution of y. The difference is that at least one of the parameters of the distribution is the unknown variable instead of y. As one example, which is seen commonly in the subsequent sections, if we know the random variable Y is some random function $g(X)$ plus a Gaussian distributed random variable such that

$$Y = g(X) + v, \qquad v \sim N[0, V], \tag{4.60}$$

then the likelihood function is written as a Gaussian distribution with a mean of $g(x)$ and variance of V as,

$$\mathcal{L}(y|x) = \frac{1}{\sqrt{2\pi V}} e^{-(y-g(x))^2/2V}. \tag{4.61}$$

4.5.2.3 The Filtering Estimation Derivation for The Local Level DLM

Before we give the Recursive Bayesian Estimation derivation for the filtering recurrence relations of the general DLM, we show the same process for the filtering recurrence relations of the local level DLM in this section. This makes for a nicer and simpler introduction with the benefit of how to extend the process to the general DLM afterward being very obvious.

The observation and evolution equations for the local level DLM are given here again for convenience as,

$$Y_i = M_i + v_i, \qquad v_i \sim N[0, V_i] \qquad i = 1, 2, ..., n, \tag{4.62}$$

$$M_i = M_{i-1} + w_i, \qquad w_i \sim N[0, W_i] \qquad i = 1, 2, ..., n. \tag{4.63}$$

We define values of the random variable M_i as μ_i and the initial conditions (mean and variance of M_0) to be μ_0^0 and P_0^0. Furthermore, assume the random variable M_0 is Gaussian distributed so that $p(\mu_0) \sim N[\mu_0^0, P_0^0]$. The objective is to find the mean μ_i^i and variance P_i^i

of the conditional distribution $p(\mu_i|\tilde{y}_i)$ for any i. First, notice that for $i = 1$, using Bayes theorem, we may write,

$$p(\mu_1|y_1) \propto \mathcal{L}(y_1|\mu_1)p(\mu_1). \tag{4.64}$$

Then, for $i = 2$ we may write similarly that,

$$p(\mu_2|y_2, y_1) \propto \mathcal{L}(y_2|\mu_2, y_1)p(\mu_2|y_1), \tag{4.65}$$

and for general i we have,

$$p(\mu_i|\tilde{y}_i) \propto \mathcal{L}(y_i|\mu_i, \tilde{y}_{i-1})p(\mu_i|\tilde{y}_{i-1}), \tag{4.66}$$

where we use the $\tilde{y}_{i-1} = y_1, y_2, ..., y_{i-1}$ notation.

The likelihood $\mathcal{L}(y_i|\mu_i, \tilde{y}_{i-1})$ is given as described in Section 4.5.2.2 in light of the observation equation (Equation 4.62). It is Gaussian with mean written as μ_i and variance V_i. This likelihood does not depend on any y_s for $s < i$ when μ_i is given (we see this because Y_i is completely described by the observation equation when μ_i is given). So, we can technically write the likelihood as $\mathcal{L}(y_i|\mu_i)$ without the conditioning on $\tilde{y}_{i-1}$. Putting this into Equation 4.66 we have,

$$p(\mu_i|\tilde{y}_i) \propto \mathcal{L}(y_i|\mu_i)p(\mu_i|\tilde{y}_{i-1}), \tag{4.67}$$

for general i.

The prior distribution $p(\mu_i|\tilde{y}_{i-1})$ is Gaussian distributed and the posterior distribution $p(\mu_i|\tilde{y}_i)$ is also Gaussian distributed. To prove that this is true for any i we use the following argument: the first prior $p(\mu_1)$ is clearly Gaussian distributed by looking at the evolution equation with $i = 1$, since M_0 is Gaussian and linear combinations of Gaussian random variables are also Gaussian. Now, Appendix N shows that for Bayes theorem a Gaussian prior and a Gaussian likelihood result in a Gaussian posterior. Therefore, the first posterior $p(\mu_1|y_1)$ must also be Gaussian. With this we can show then that the next prior $p(\mu_2|y_1)$ is Gaussian because the random variable that represents this distribution (define it as M_2^1) is described by the random variable that describes the first posterior $p(\mu_1|y_1)$ (define as M_1^1)

with the evolution equation such that $M_2^1 = M_1^1 + w_2$, and linear combinations of Gaussian random variables are Gaussian random variables. So, this means that the prior $p(\mu_2|y_1)$ is Gaussian and therefore, again, by the result of Appendix N the next posterior $p(\mu_2|y_2)$ is also Gaussian. This can be continued for all i.

Appendix N also shows the resulting mean and variance of the posterior distribution for when the prior and likelihood are Gaussian with known means and variances. They are given by Equations N.8 and N.16 in the appendix. We have stated the expectation and variance of the likelihood already as μ_i and V_i, and we can find the expectation and variance of the prior distribution $p(\mu_i|\tilde{y}_{i-1})$ as follows:

$$\begin{aligned} \mathrm{E}[M_i|\tilde{y}_{i-1}] &= \mathrm{E}[M_{i-1} + w_i|\tilde{y}_{i-1}] && (4.68)\\ &= \mathrm{E}[M_{i-1}|\tilde{y}_{i-1}] && (4.69)\\ &= \mu_{i-1}^{i-1}, && (4.70) \end{aligned}$$

and,

$$\begin{aligned} \mathrm{Var}[M_i|\tilde{y}_{i-1}] &= \mathrm{Var}[M_{i-1} + w_i|\tilde{y}_{i-1}] && (4.71)\\ &= \mathrm{Var}[M_{i-1}|\tilde{y}_{i-1}] + \mathrm{Var}[w_i] && (4.72)\\ &= P_{i-1}^{i-1} + W_i. && (4.73) \end{aligned}$$

So, using the equations in the appendix we see that the mean and variance of the posterior distribution $p(\mu_i|\tilde{y}_i)$ are given as,

$$\mu_i^i = \mu_{i-1}^{i-1} + \frac{P_{i-1}^{i-1} + W_i}{P_{i-1}^{i-1} + W_i + V_i}(y_i - \mu_{i-1}^{i-1}) \quad (4.74)$$

$$P_i^i = \frac{P_{i-1}^{i-1} + W_i}{P_{i-1}^{i-1} + W_i + V_i} V_i. \quad (4.75)$$

These are the filtering recurrence relations for the local level DLM. It can easily be verified that they are consistent with the recurrence relations for the general DLM filtering problem given in Section 4.4.2 by setting $\mathbf{F}_i$, $\mathbf{G}_i$, $\mathbf{W}_i$, and $\mathbf{V}_i$ equal to the local level DLM's model matrices of 1, 1, W_i, and V_i respectively.

4.5.2.4 The Filtering Estimation Derivation

Finally, we show in this section the Recursive Bayesian Estimation derivation for the filtering recurrence relations of the general DLM. We follow essentially the same process as what was shown in the previous section for the local level DLM case.

The general DLM evolution and observations equations stated again for convenience are given as,

$$\mathbf{Y}_i = \mathbf{F}_i\mathbf{X}_i + \mathbf{v}_i, \qquad \mathbf{v}_i \sim N[0, \mathbf{V}_i] \qquad i = 1, 2, ..., n, \tag{4.76}$$

$$\mathbf{X}_i = \mathbf{G}_i\mathbf{X}_{i-1} + \mathbf{w}_i, \qquad \mathbf{w}_i \sim N[0, \mathbf{W}_i] \qquad i = 1, 2, ..., n. \tag{4.77}$$

The initial conditions (mean and variance of $\mathbf{X}_0$) are $\mathbf{x}_0^0$ and $\mathbf{P}_0^0$, with $p(\mathbf{x}_0) \sim N[\mathbf{x}_0, \mathbf{P}_0]$. The objective is to find the mean $\mathbf{x}_i^i$ and variance $\mathbf{P}_i^i$ of the conditional distribution $p(\mathbf{x}_i|\tilde{\mathbf{y}}_i)$ for any i. Like in Equation 4.67 for the local level DLM case, we may write,

$$p(\mathbf{x}_i|\tilde{\mathbf{y}}_i) \propto \mathcal{L}(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{x}_i|\tilde{\mathbf{y}}_{i-1}), \tag{4.78}$$

using Bayes theorem. The likelihood $\mathcal{L}(\mathbf{y}_i|\mathbf{x}_i)$ is given as described in Section 4.5.2.2 in light of the observation equation (Equation 4.76) (i.e. Gaussian with mean written as $\mathbf{F}_i\mathbf{X}_i$ and variance $\mathbf{V}_i$). The prior $p(\mathbf{x}_i|\tilde{\mathbf{y}}_{i-1})$ and posterior $p(\mathbf{x}_i|\tilde{\mathbf{y}}_i)$ are Gaussian distributed for all i with the same argument given in the previous section. The variables and evolution equation are slightly different here, but otherwise the argument does not change. The prior has expectation

$$\mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{y}}_{i-1}] = \mathbf{x}_i^{i-1}, \tag{4.79}$$

and covariance

$$\mathrm{Cov}[\mathbf{X}_i|\tilde{\mathbf{y}}_{i-1}] = \mathbf{P}_i^{i-1}, \tag{4.80}$$

simply by definition.

Now, Appendix N, in addition to the univariate case used in the last section, also shows the resulting mean and covariance matrix of the posterior distribution for the multivariate

case when the likelihood and prior are Gaussian distributed. These are given by Equations N.37 and N.28 in the appendix. Using these equations for the problem at hand we see that the mean and covariance of the posterior $p(\mathbf{x}_i|\tilde{\mathbf{y}}_i)$ are given as,

$$\mathbf{x}_i^i = \mathbf{x}_i^{i-1} + \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{V}_i + \mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T)^{-1}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i^{i-1}) \tag{4.81}$$

$$\mathbf{P}_i^i = \mathbf{P}_i^{i-1} - \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{V}_i + \mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T)^{-1}\mathbf{F}_i\mathbf{P}_i^{i-1}. \tag{4.82}$$

These are the filtering recurrence relations for the general DLM, consistent with what has already been given in Section 4.4.2. So, we have seen our first derivation for the filtering recurrence relations.

4.5.2.5 A Note on Conditioning with a Random Variable

Before we use a Recursive Bayesian Estimation like approach to obtain the smoothing recurrence relations, we have a short discussion in this section about a topic that is seen in this approach. The topic is probability distributions that are conditioned on random variables, rather than numbers. Consider a Gaussian distributed conditional probability distribution $p(x|y)$ with mean as some function $f(y)$ and variance V, written as,

$$p(x|y) = \frac{1}{\sqrt{2\pi V}}e^{-(x-f(y))^2/2V}. \tag{4.83}$$

We may also define the density $p(x|Y)$ as,

$$p(x|Y) = \frac{1}{\sqrt{2\pi V}}e^{-(x-f(Y))^2/2V}. \tag{4.84}$$

What perhaps seems odd about this is that the density $p(x|Y)$, expectation $\mathrm{E}[X|Y]$, and variance $\mathrm{Var}[X|Y]$ are random variables themselves. This is not unfounded and can be a common formulation. The law of total expectation can be used in such scenarios to find the "total" expectation of the random variable X. This expectation, written as $\mathrm{E}[X]$, can be found by evaluating the law of total expectation formula, given as,

$$\mathrm{E}[X] = \mathrm{E}[\mathrm{E}[X|Y]]. \tag{4.85}$$

Similarly, there is a law of total variance that states that the "total" variance $\text{Var}[X]$ can be found with the following formula:

$$\text{Var}[Y] = \text{E}[\text{Var}[X|Y]] + \text{Var}[\text{E}[X|Y]]. \tag{4.86}$$

To generalize these from random variables to random vectors $\mathbf{X}$ and $\mathbf{Y}$, the law of total expectation is pretty simple with $\text{E}[\mathbf{X}] = \text{E}[\text{E}[\mathbf{X}|\mathbf{Y}]]$, and the law of total covariance (using covariance matrices) becomes

$$\text{Cov}[\mathbf{Y}] = \text{E}[\text{Cov}[\mathbf{X}|\mathbf{Y}]] + \text{Cov}[\text{E}[\mathbf{X}|\mathbf{Y}]]. \tag{4.87}$$

Lastly, we note that the likelihood from Section 4.5.2.2 could similarly be written with a random variable Y, instead of the number y, as,

$$\mathcal{L}(Y|x) = \frac{1}{\sqrt{2\pi V}} e^{-(Y-g(x))^2/2V}. \tag{4.88}$$

4.5.2.6 The Smoothing Estimation Derivation

The derivation for the smoothing recurrence relations using Bayes theorem is a little bit less intuitive and more tricky than the filtering derivation just given. The objective is to calculate $\text{E}[\mathbf{X}_i|\tilde{\mathbf{y}}_n]$ and $\text{Cov}[\mathbf{X}_i|\tilde{\mathbf{y}}_n]$ ($\mathbf{x}_i^n$ and $\mathbf{P}_i^n$) where $i < n$. By the law of total expectation and the law of total covariance, we see that we can calculate these from the quantities $\mathbf{s}_i^n = \text{E}[\mathbf{X}_i|\mathbf{X}_{i+1}, \tilde{\mathbf{y}}_n]$ and $\mathbf{S}_i^n = \text{Cov}[\mathbf{X}_i|\mathbf{X}_{i+1}, \tilde{\mathbf{y}}_n]$ with the following two equations:

$$\text{E}[\mathbf{X}_i|\tilde{\mathbf{y}}_n] = \text{E}[\mathbf{s}_i^n|\tilde{\mathbf{y}}_n] \tag{4.89}$$

$$\text{Cov}[\mathbf{X}_i|\tilde{\mathbf{y}}_n] = \text{E}[\mathbf{S}_i^n|\tilde{\mathbf{y}}_n] + \text{Cov}[\mathbf{s}_i^n|\tilde{\mathbf{y}}_n]. \tag{4.90}$$

The variables $\mathbf{s}_i^n$ and $\mathbf{S}_i^n$ have been defined here for convenience. So, the approach we will take in this derivation is to first try to find $p(\mathbf{x}_i|\mathbf{X}_{i+1}, \tilde{\mathbf{y}}_n)$ so that the above equations can then be used to find $p(\mathbf{x}_i|\tilde{\mathbf{y}}_n)$.

The first thing we notice is that we can write,

$$p(\mathbf{x}_i|\mathbf{X}_{i+1}, \tilde{\mathbf{y}}_n) = p(\mathbf{x}_i|\mathbf{X}_{i+1}, \tilde{\mathbf{y}}_i), \tag{4.91}$$

because $\mathbf{X}_i$ and $\mathbf{Y}_s$ for any s such that $s > i$ are conditionally independent given $\mathbf{X}_{i+1}$ (see Appendix O). Now, using Bayes theorem we may write,

$$p(\mathbf{x}_i|\mathbf{X}_{i+1}) \propto \mathcal{L}(\mathbf{X}_{i+1}|\mathbf{x}_i)p(\mathbf{x}_i), \tag{4.92}$$

and then conditioning on $\tilde{\mathbf{y}}_i$ we may write $p(\mathbf{x}_i|\mathbf{X}_{i+1}, \tilde{\mathbf{y}}_i)$, the distribution of interest, as,

$$p(\mathbf{x}_i|\mathbf{X}_{i+1}, \tilde{\mathbf{y}}_i) \propto \mathcal{L}(\mathbf{X}_{i+1}|\mathbf{x}_i, \tilde{\mathbf{y}}_i)p(\mathbf{x}_i, \tilde{\mathbf{y}}_i). \tag{4.93}$$

We see in this equation that the prior is the filtering estimation distribution, so it has a mean $\mathbf{x}_i^i$ and covariance $\mathbf{P}_i^i$, and that the likelihood is given as described in Section 4.5.2.2 in light of the evolution equation. This likelihood does not depend on y_i for any i (this is seen clearly by the evolution equation). So, we may write the likelihood simply as $\mathcal{L}(\mathbf{X}_{i+1}|\mathbf{x}_i)$, and its mean and covariance are $\mathbf{G}_{i+1}\mathbf{X}_i$ and $\mathbf{W}_{i+1}$ respectively. With this information the mean and covariance of the posterior distribution can be found by using Equations N.37 and N.28 in Appendix N. The results are:

$$\begin{aligned}
\mathbf{s}_i &= \mathbf{x}_i^i + \mathbf{P}_i^i\mathbf{G}_{i+1}^T(\mathbf{W}_{i+1} + \mathbf{G}_{i+1}\mathbf{P}_i^i\mathbf{G}_{i+1}^T)^{-1}(\mathbf{X}_{i+1} - \mathbf{G}_{i+1}\mathbf{x}_i^i) && (4.94)\\
&= \mathbf{x}_i^i + \mathbf{P}_i^i\mathbf{G}_{i+1}^T{\mathbf{P}_{i+1}^i}^{-1}(\mathbf{X}_{i+1} - \mathbf{x}_{i+1}^i) && (4.95)\\
\mathbf{S}_i &= \mathbf{P}_i^i - \mathbf{P}_i^i\mathbf{G}_{i+1}^T(\mathbf{W}_{i+1} + \mathbf{G}_{i+1}\mathbf{P}_i^i\mathbf{G}_{i+1}^T)^{-1}\mathbf{G}_{i+1}\mathbf{P}_i^i && (4.96)\\
&= \mathbf{P}_i^i - \mathbf{P}_i^i\mathbf{G}_{i+1}^T{\mathbf{P}_{i+1}^i}^{-1}\mathbf{G}_{i+1}\mathbf{P}_i^i, && (4.97)
\end{aligned}$$

where the superscript n for $\mathbf{s}_i^n$ and $\mathbf{S}_i^n$ have been dropped because they are redundant in light of Equation 4.91.

The quantities in Equations 4.89 and 4.90 can now be calculated as,

$$\begin{aligned}
\mathrm{E}[\mathbf{s}_i|\tilde{\mathbf{y}}_n] &= \mathbf{x}_i^i + \mathbf{P}_i^i\mathbf{G}_{i+1}^T{\mathbf{P}_{i+1}^i}^{-1}(\mathbf{x}_{i+1}^n - \mathbf{x}_{i+1}^i) && (4.98)\\
\mathrm{E}[\mathbf{S}_i|\tilde{\mathbf{y}}_n] &= \mathbf{P}_i^i - \mathbf{P}_i^i\mathbf{G}_{i+1}^T{\mathbf{P}_{i+1}^i}^{-1}\mathbf{G}_{i+1}\mathbf{P}_i^i && (4.99)\\
\mathrm{Cov}[\mathbf{s}_i|\tilde{\mathbf{y}}_n] &= \mathbf{P}_i^i\mathbf{G}_{i+1}^T{\mathbf{P}_{i+1}^i}^{-1}\mathbf{P}_{i+1}^n{\mathbf{P}_{i+1}^i}^{-1}\mathbf{G}_{i+1}\mathbf{P}_i^i, && (4.100)
\end{aligned}$$

by recalling that by definition $\mathrm{E}[\mathbf{X}_{i+1}|\tilde{\mathbf{y}}_n] = \mathbf{x}_{i+1}^n$ and $\mathrm{Var}[\mathbf{X}_{i+1}|\tilde{\mathbf{y}}_n] = \mathbf{P}_{i+1}^n$. So, we can finally put these into Equations 4.89 and 4.90 to calculate the mean and covariance of the distribution we are looking for of $p(\mathbf{x}_i|\tilde{\mathbf{y}}_n)$. They are found to be:

$$\mathbf{x}_i^n = \mathbf{x}_i^i + \mathbf{P}_i^i \mathbf{G}_{i+1}^T {\mathbf{P}_{i+1}^i}^{-1} (\mathbf{x}_{i+1}^n - \mathbf{x}_{i+1}^i) \tag{4.101}$$

$$\mathbf{P}_i^n = \mathbf{P}_i^i + \mathbf{P}_i^i \mathbf{G}_{i+1}^T {\mathbf{P}_{i+1}^i}^{-1} (\mathbf{P}_{i+1}^n - \mathbf{P}_{i+1}^i) {\mathbf{P}_{i+1}^i}^{-1} \mathbf{G}_{i+1} \mathbf{P}_i^i. \tag{4.102}$$

These are the smoothing recurrence relations for the general DLM, consistent with what has already been given in Section 4.4.3.

Lastly, we note that we have not been completely rigorous in this derivation yet having not made any argument for the smoothing distribution $p(\mathbf{x}_i|\tilde{\mathbf{y}}_n)$ being Gaussian. The distribution $p(\mathbf{x}_i|\mathbf{X}_{i+1}, \tilde{\mathbf{y}}_i)$ is the posterior of a Gaussian likelihood and a Gaussian prior (Equation 4.93) and is therefore also Gaussian. But, the distribution is itself a random vector. Its mean is the Gaussian random vector $\mathbf{s}_i$, and $\mathbf{s}_i$ is a linear function of the Gaussian random vector $\mathbf{X}_{i+1}$, so it is Gaussian. Since $p(\mathbf{x}_i|\mathbf{X}_{i+1}, \tilde{\mathbf{y}}_i)$ is written as a Gaussian distribution with mean as a Gaussian random vector, the distribution $p(\mathbf{x}_i|\tilde{\mathbf{y}}_n)$, with $\mathbf{X}_{i+1}$ absorbed in, is also a Gaussian distribution. This can be verified and is not shown here in this thesis. So, we have seen our first derivation for the smoothing recurrence relations.

4.5.3 Minimum Mean Squared Error

The original derivation used by Kalman in 1960 for the filtering recurrence relations is based on the minimum mean squared error (MMSE) statistic. An introduction to the MMSE statistic is given in Appendix C. For the general DLM, the MSE of an estimator $\mathbf{A}_i$ of the random vector $\mathbf{X}_i$ is given by,

$$MSE = \mathrm{E}[(\mathbf{A}_i - \mathbf{X}_i)^T (\mathbf{A}_i - \mathbf{X}_i))] \tag{4.103}$$

$$= \mathrm{tr}(\mathrm{E}[(\mathbf{A}_i - \mathbf{X}_i)(\mathbf{A}_i - \mathbf{X}_i)^T]). \tag{4.104}$$

If the data $\tilde{\mathbf{y}}_n$ is observed and the estimate $\mathbf{a}_i$ is some function of $\tilde{\mathbf{y}}_n$ (and its corresponding estimator $\mathbf{A}_i$ is the same function of $\tilde{\mathbf{Y}}_n$), then the MSE of the estimate $\mathbf{a}_i$ can be written

as,

$$MSE = \mathrm{E}[(\mathbf{a}_i - \mathbf{X}_i)^T(\mathbf{a}_i - \mathbf{X}_i))|\tilde{\mathbf{y}}_n]. \tag{4.105}$$

In Section 4.5.3.1 it is shown that the MMSE estimate $\mathbf{a}_{i,m}$ out of all possible estimates for $\mathbf{a}_i$ is given as,

$$\mathbf{a}_{i,m} = \mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{y}}_n], \tag{4.106}$$

which is the conditional expectation we defined as $\mathbf{x}_i^n$ before. In Section 4.5.3.2 what is called the "error covariance matrix" is discussed and in Section 4.5.3.3 the MMSE derivation for the filtering problem is carried out. Also, in Section 4.5.4 a difficult proof made by Kalman in his seminal paper which is needed for part of the derivation in Section 4.5.3.3 is given. We make note of this in Section 4.5.3.3 when it comes up. This point is often overlooked by other resources that show this MMSE derivation for the Kalman filter.

4.5.3.1 MMSE as a Conditional Expectation

Given that the estimate we have labelled as $\mathbf{a}_i$ is some function of the observed data $g(\tilde{\mathbf{y}}_n)$, the MMSE problem we have set up is the same as the one described in Section C.1 of Appendix C. Using the result of this section we have,

$$\mathbf{a}_{i,m} = \mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{y}}_n] \tag{4.107}$$

$$= \mathbf{x}_i^n. \tag{4.108}$$

So, the $\mathbf{a}_{i,m}$ and $\mathbf{A}_{i,m}$ notation will be dropped now and the MMSE estimate is written as $\mathbf{x}_i^n$, and its corresponding estimator as $\mathbf{X}_i^n$ moving forward. Also, any estimate of the state vector (not necessarily MMSE) is written as $\mathbf{x}_i^{n\prime}$ and its corresponding estimator as $\mathbf{X}_i^{n\prime}$.

Plugging in $\tilde{\mathbf{Y}}_n$ for $\tilde{\mathbf{y}}_n$ we have the MMSE estimator given as,

$$\mathbf{X}_i^n = \mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{Y}}_n]. \tag{4.109}$$

Taking the expected value of this random vector gives $\mathrm{E}[\mathbf{X}_i^n] = \mathrm{E}[\mathbf{X}_i]$ by the law of total expectation. This shows that the estimator is unbiased. In fact, these expectations are also

usually equal to zero when the prior expectation $\mathrm{E}[\mathbf{X}_0]$ is chosen to be zero (see Appendix M).

4.5.3.2 The Error Covariance Matrix

The matrix

$$\mathbf{P}_i^{n\prime} = \mathrm{E}[(\mathbf{X}_i^{n\prime} - \mathbf{X}_i)(\mathbf{X}_i^{n\prime} - \mathbf{X}_i)^T] \tag{4.110}$$

is the matrix where if $\mathbf{X}_i^{n\prime}$ is selected so the trace is minimized, then the selection is the MMSE estimate $\mathbf{X}_i^n$. We have denoted this matrix by $\mathbf{P}_i^{n\prime}$. When the MMSE estimator $\mathbf{X}_i^n$ is used we will denote the matrix as $\mathbf{P}_i^n$, which is what we call the error covariance matrix. So, the error covariance matrix is given as,

$$\mathbf{P}_i^n = \mathrm{E}[(\mathbf{X}_i^n - \mathbf{X}_i)(\mathbf{X}_i^n - \mathbf{X}_i)^T]. \tag{4.111}$$

This error covariance matrix $\mathbf{P}_i^n$ should not be confused with the interpretation of $\mathbf{P}_i^n$ presented in the Recursive Bayesian Estimation derivations, even though we will see that it ends up as the same quantity here. We note that we refer to this as the error covariance matrix because we have,

$$\begin{aligned}
\mathbf{P}_i^n &= \mathrm{E}[(\mathbf{X}_i^n - \mathbf{X}_i)(\mathbf{X}_i^n - \mathbf{X}_i)^T] && (4.112)\\
&= \mathrm{Cov}[\mathbf{X}_i^n - \mathbf{X}_i] - \mathrm{E}[\mathbf{X}_i^n - \mathbf{X}_i]\mathrm{E}[\mathbf{X}_i^n - \mathbf{X}_i]^T && (4.113)\\
&= \mathrm{Cov}[\mathbf{X}_i^n - \mathbf{X}_i], && (4.114)
\end{aligned}$$

since the expectation of the "error" is zero (i.e. $\mathrm{E}[\mathbf{X}_i^n - \mathbf{X}_i] = 0$).

4.5.3.3 The Filtering Estimation Derivation

In this section, the derivation of the filtering recurrence relations for the DLM is given with the MMSE criteria. As a reference to this, we cite (Brown and Hwang, 2012). For this derivation, we seek to find an MMSE estimator of the state vector at index i using all $\tilde{\mathbf{Y}}_i$ (what we have defined as $\mathbf{X}_i^i$) given that the MMSE estimator of $\mathbf{X}_{i-1}$ using all $\tilde{\mathbf{Y}}_{i-1}$ is known (what we have defined as $\mathbf{X}_{i-1}^{i-1}$).

With $\mathbf{X}_{i-1}^{i-1}$ known and $\mathbf{X}_i^n = \mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{Y}}_n]$ being true for the MMSE estimator, we have the MMSE estimator of the state represented by $\mathbf{X}_i$ using all $\tilde{\mathbf{Y}}_{i-1}$ given as (see Section 4.5.1),

$$\mathbf{X}_i^{i-1} = \mathbf{G}_i \mathbf{X}_{i-1}^{i-1}. \tag{4.115}$$

Given this one-step-ahead $\mathbf{X}_i^{i-1}$ estimator, we now seek a better estimator of the state vector at index i by using the additional piece of information, $\mathbf{Y}_i$. The equation that we use to do this is arbitrarily defined (at least consider it to be arbitrary for now, until the reasoning is shown in Section 4.5.4) to be in the following form:

$$\mathbf{X}_i^{i'} = \mathbf{X}_i^{i-1} + \mathbf{K}_i'(\mathbf{Y}_i - \mathbf{F}_i \mathbf{X}_i^{i-1}), \tag{4.116}$$

where $\mathbf{K}_i'$ is any matrix and $\mathbf{X}_i^{i'}$ is the estimator of the state vector at index i, where again the prime denotes that it is any estimator, not necessarily the MMSE estimator. This equation is commonly called the "update" equation by Kalman filter practitioners. It "updates" the prediction estimate $\mathbf{X}_i^{i-1}$ of the state vector given one piece of additional data $\mathbf{Y}_i$, where essentially, the matrix $\mathbf{K}_i'$ is related to how much weight the new $\mathbf{Y}_i$ carries in the estimation. Again, the justification for the form of Equation 4.116 is given in Section 4.5.4. Some authors reference the Recursive Bayesian Estimation derivations we have shown as justification for this form, but Kalman deduced the form without the use of Bayes theorem in his seminal paper. The justification that Kalman used is what we give in Section 4.5.4.

Now, the MMSE estimator $\mathbf{X}_i^i$ will be found by finding the matrix $\mathbf{K}_i'$ that minimizes the MSE. We denote this optimal matrix as $\mathbf{K}_i$ without the prime as well. First, we will evaluate $\mathbf{P}_i^{i'}$ so that we can then minimize its trace to find $\mathbf{K}_i$. By inserting Equation 4.116 into the definition of $\mathbf{P}_i^{i'}$ we have,

$$\mathbf{P}_i^{i'} = \mathrm{Cov}[\mathbf{X}_i^{i'} - \mathbf{X}_i] \tag{4.117}$$

$$= \mathrm{Cov}[\mathbf{X}_i^{i-1} + \mathbf{K}_i'(\mathbf{Y}_i - \mathbf{F}_i \mathbf{X}_i^{i-1}) - \mathbf{X}_i]. \tag{4.118}$$

Then, using the observation equation for $\mathbf{Y}_i$ we have,

$$\mathbf{P}_i^{i'} = \text{Cov}[\mathbf{X}_i^{i-1} + \mathbf{K}_i'(\mathbf{F}_i\mathbf{X}_i + \mathbf{v}_i - \mathbf{F}_i\mathbf{X}_i^{i-1}) - \mathbf{X}_i] \tag{4.119}$$

$$= \text{Cov}[(\mathbf{I} - \mathbf{K}_i'\mathbf{F}_i)(\mathbf{X}_i^{i-1} - \mathbf{X}_i) + \mathbf{K}_i'\mathbf{v}_i]. \tag{4.120}$$

Since $\mathbf{v}_i$ is uncorrelated to both $\mathbf{X}_i^{i-1}$ and $\mathbf{X}_i$ (for $\mathbf{X}_i^{i-1}$ the data $\mathbf{y}_i$ has not entered the picture yet, so therefore neither has $\mathbf{v}_i$, and for $\mathbf{X}_i$ recall the evolution equation and that $\mathbf{w}_i$ and $\mathbf{v}_i$ have been defined to be uncorrelated) we have,

$$\mathbf{P}_i^{i'} = \text{Cov}[(\mathbf{I} - \mathbf{K}_i'\mathbf{F}_i)(\mathbf{X}_i^{i-1} - \mathbf{X}_i)] + \text{Cov}[\mathbf{K}_i'\mathbf{v}_i] \tag{4.121}$$

$$= (\mathbf{I} - \mathbf{K}_i'\mathbf{F}_i)\text{Cov}[\mathbf{X}_i^{i-1} - \mathbf{X}_i](\mathbf{I} - \mathbf{K}_i'\mathbf{F}_i)^T + \mathbf{K}_i'\mathbf{V}_i{\mathbf{K}_i'}^T \tag{4.122}$$

$$= (\mathbf{I} - \mathbf{K}_i'\mathbf{F}_i)\mathbf{P}_i^{i-1}(\mathbf{I} - \mathbf{K}_i'\mathbf{F}_i)^T + \mathbf{K}_i'\mathbf{V}_i{\mathbf{K}_i'}^T. \tag{4.123}$$

This will now be minimized with respect to $\mathbf{K}_i'$. We rewrite it in the following form first to make the derivatives easier:

$$\mathbf{P}_i^{i'} = \mathbf{P}_i^{i-1} - \mathbf{K}_i'\mathbf{F}_i\mathbf{P}_i^{i-1} - \mathbf{P}_i^{i-1}\mathbf{F}_i^T{\mathbf{K}_i'}^T + \mathbf{K}_i'(\mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T + \mathbf{V}_i){\mathbf{K}_i'}^T. \tag{4.124}$$

Using the identities in Appendix P for the derivative of traces of matrices with respect to a matrix we have,

$$\frac{d\,\text{tr}(\mathbf{P}_i^{i'})}{d\mathbf{K}_i'} = -(\mathbf{F}_i\mathbf{P}_i^{i-1})^T - (\mathbf{F}_i\mathbf{P}_i^{i-1})^T + 2\mathbf{K}_i(\mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T + \mathbf{V}_i) = 0. \tag{4.125}$$

Therefore we find the matrix $\mathbf{K}_i$ to be given as,

$$\mathbf{K}_i = \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T + \mathbf{V}_i)^{-1}. \tag{4.126}$$

So, the MMSE estimator $\mathbf{X}_i^i$ is given by using this as the $\mathbf{K}_i'$ matrix in Equation 4.116 and its corresponding estimate is given as,

$$\mathbf{x}_i^i = \mathbf{x}_i^{i-1} + \mathbf{K}_i(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i^{i-1}). \tag{4.127}$$

This is the same form of the recurrence relation for filtering estimation given and derived in previous sections.

We would now also like to use this optimal $\mathbf{K}_i$ to find the recurrence relation for the error covariance matrix $\mathbf{P}_i^i$ of the MMSE estimate. Using $\mathbf{K}_i$ as the $\mathbf{K}_i'$ matrix in Equation 4.124 we find that,

$$\begin{aligned}\mathbf{P}_i^i &= \mathbf{P}_i^{i-1} - \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T + \mathbf{V}_i)^{-1}\mathbf{F}_i\mathbf{P}_i^{i-1} - \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T + \mathbf{V}_i)^{-1}\mathbf{F}_i\mathbf{P}_i^{i-1} \\ &\quad + \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T + \mathbf{V}_i)^{-1}(\mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T + \mathbf{V}_i)(\mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T + \mathbf{V}_i)^{-1}\mathbf{F}_i\mathbf{P}_i^{i-1} \quad (4.128) \\ &= \mathbf{P}_i^{i-1} - \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T + \mathbf{V}_i)^{-1}\mathbf{F}_i\mathbf{P}_i^{i-1} \quad (4.129) \\ &= \mathbf{P}_i^{i-1} - \mathbf{K}_i\mathbf{F}_i\mathbf{P}_i^{i-1}. \quad (4.130)\end{aligned}$$

This is the same form of the filtering recurrence relation given for $\mathbf{P}_i^i$ in previous sections, and also the same as the quantity $\text{Cov}[\mathbf{X}_i|\tilde{\mathbf{y}}_i]$, which we also defined as $\mathbf{P}_i^i$, found with the recursive application of Bayes theorem.

However, if we are looking at this MMSE derivation stand-alone we have yet to specify what $\mathbf{P}_i^{i-1}$ is. Because we have actually only used it as a definition in the process carried out so far. Evaluating it explicitly we have,

$$\begin{aligned}\mathbf{P}_i^{i-1} &= \text{Cov}[\mathbf{X}_i^{i-1} - \mathbf{X}_i] \quad (4.131) \\ &= \text{Cov}[\mathbf{G}_i\mathbf{X}_{i-1}^{i-1} - \mathbf{G}_i\mathbf{X}_{i-1} - \mathbf{w}_i]. \quad (4.132)\end{aligned}$$

Since $\mathbf{w}_i$ is uncorrelated to both $\mathbf{X}_{i-1}^{i-1}$ and $\mathbf{X}_{i-1}$ we have,

$$\begin{aligned}\mathbf{P}_i^{i-1} &= \text{Cov}[\mathbf{G}_i(\mathbf{X}_{i-1}^{i-1} - \mathbf{X}_{i-1})] + \text{Cov}[\mathbf{w}_i] \quad (4.133) \\ &= \mathbf{G}_i\mathbf{P}_{i-1}^{i-1}\mathbf{G}_i^T + \mathbf{W}_i. \quad (4.134)\end{aligned}$$

So finally we have now shown that the recurrence relations for both $\mathbf{x}_i^i$ and $\mathbf{P}_i^i$ obtained in this section are the same as those given and argued in previous sections.

Lastly, we make the note that if someone wished to use a matrix $\mathbf{K}_i'$ that is not MMSE, then they could do so and they would use Equation 4.123 for updating $\mathbf{P}_i^i$ instead of Equation 4.130. This is called the Joseph form of the error covariance matrix. In the next section, we will give Kalman's justification for the form of the "update" equation.

4.5.4 Kalman's Justification

In this section, we justify the form of Equation 4.116, as done in the seminal work of (Kalman, 1960). We also cite (Jazwinski, 1970), who provides a more throughout explanation of this content that the work presented here follows more closely.

To proceed with the justification, we must first develop the concepts of orthogonality and orthogonal projections of random variables and random vectors. These concepts are analogous to orthogonality concepts in linear algebra. Consider the random variables $Y_1, Y_2, ..., Y_n$. The set of all linear combinations of these random variables is said to form a vector space, in analogy to the way a linear combination of vectors forms a vector space. The set of all linear combinations of these random variables can be written as,

$$\sum_{i=0}^{n} a_i Y_i, \tag{4.135}$$

where the coefficients a_i can take on any values. The vector space this forms will be called the Υ_n vector space or simply Υ_n space.

The definition of orthogonality of random variables is given as follows: two random variables U and V are said to be orthogonal if $\mathrm{E}[UV] = 0$. A trick commonly used in this section is to write random variables as two components, one orthogonal to a given vector space and one in the vector space. Consider the random variable S:

$$S = \bar{S} + \tilde{S}, \tag{4.136}$$

where we define $\tilde{S}$ to be the part of S orthogonal to the chosen vector space and $\bar{S}$ as the component of S in the chosen vector space. We call $\bar{S}$ the "orthogonal projection" of S onto the chosen vector space.

For random vectors, we will develop a story that is pretty much the same. Consider the random vectors $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n$ where each is of length m. The set of all linear combinations of the random variables that make up these random vectors is said to form a vector space. This set can be written as,

$$\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}Y_{ij}, \tag{4.137}$$

where Y_{ij} is the jth component of $\mathbf{Y}_i$ and the coefficients a_{ij} can take on any values. The vector space this forms will be called $\boldsymbol{\Upsilon}_n$ space.

Two random vectors $\mathbf{U}$ and $\mathbf{V}$ are said to be orthogonal if each of their components are orthogonal to each other. Mathematically, if $\mathbf{U}$ and $\mathbf{V}$ are the same length this means that $\mathrm{E}[\mathbf{U}\mathbf{V}^T] = \mathbf{0}$.

Similarly to the trick above but for random vectors, we can break random vectors into two components. In fact, this is only the trick of Equation 4.136 applied for each of the elements of the random vector. Consider the random vector $\mathbf{S}$:

$$\mathbf{S} = \bar{\mathbf{S}} + \tilde{\mathbf{S}}, \tag{4.138}$$

where we define $\tilde{\mathbf{S}}$ to be the random vector where each of its elements are the part of each of $\mathbf{S}$'s elements that are orthogonal to the chosen vector space, and the elements of $\bar{\mathbf{S}}$ are the orthogonal projections of each of the elements of $\mathbf{S}$ onto the chosen vector space.

Now that this background material is set up, we can proceed with Kalman's justification of the form of Equation 4.116. Consider that the $\boldsymbol{\Upsilon}_n$ vector space is formed by the DLM random vectors $\mathbf{Y}_1, ..., \mathbf{Y}_n$ (what we have labelled as $\tilde{\mathbf{Y}}_n$ previously). Now, we refer to a proof in Appendix Q that shows that for Gaussian distributed $\mathbf{X}_i$ and $\mathbf{Y}_i$ (made by our assumptions of $\mathbf{v}_i$, $\mathbf{w}_i$, and $\mathbf{X}_0$ being Gaussian) the orthogonal projection of $\mathbf{X}_i$ onto the $\boldsymbol{\Upsilon}_n$ vector space is equal to $\mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{Y}}_n]$, which is the DLM MMSE estimator. The only caveat for this to be true is that we must choose the prior expectation of $\mathrm{E}[\mathbf{X}_0]$ to be zero, which is typically done in practice. This allows us to deduce the following: if we restrict the estimator to be in the $\boldsymbol{\Upsilon}_n$ vector space (which means that the estimator is a linear function of the random variables that make up the $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n$ random vectors) and then find the MMSE under this restriction, then this results in the MMSE estimate for when $\mathbf{X}_i$ and $\mathbf{Y}_i$ are Gaussian (i.e. no non-linear functions of the $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n$ random vectors can possibly have a smaller MSE, so we lose nothing by making this restriction).

So, this argument tells us that one aspect of the form of Equation 4.116 is good, that it is

a linear function of the random variables that make up $\boldsymbol{\Upsilon}_n$. But, the form of Equation 4.116 is still more particular than this. We will need to introduce the concept of an orthonormal basis for vector spaces before we proceed further.

In linear algebra, given a set of vectors that forms a vector space, an orthonormal basis can be found that spans the vector space with basis vectors $\mathbf{q}_i$ having the following property (this is done in practice with the Gram-Schmidt process):

$$\mathbf{q}_i \mathbf{q}_j^T = \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}. \tag{4.139}$$

In a similar vein, given a set of random variables that forms a vector space, an orthonormal basis can be found that spans the vector space with basis random variables u_i having the following property (done with a similar Gram-Schmidt process):

$$\mathrm{E}[u_i u_j] = \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}. \tag{4.140}$$

With this, we know that we can write any random variable $\bar{X}$ in a given vector space as a linear combination of the vector spaces basis random variables as,

$$\bar{X} = \sum_{j=1}^{n} a_j u_j. \tag{4.141}$$

The coefficients a_j can be determined by noting that

$$\mathrm{E}[\bar{X} u_j] = \mathrm{E}[(\sum_{i=1}^{n} a_i u_i) u_j] \tag{4.142}$$

$$= \sum_{i=1}^{n} a_i \mathrm{E}[u_i u_j] \tag{4.143}$$

$$= \sum_{i=1}^{n} a_i \delta_{ij} \tag{4.144}$$

$$= a_j. \tag{4.145}$$

Thus the coefficients are given by,

$$a_j = \mathrm{E}[\bar{X}u_j]. \tag{4.146}$$

Or, if X is written as the sum of a component $\bar{X}$ in the vector space and a component $\tilde{X}$ orthogonal to the space as described by the trick of Equation 4.136, the coefficients are also given as,

$$a_j = \mathrm{E}[Xu_j], \tag{4.147}$$

since we have,

$$\mathrm{E}[Xu_j] = \mathrm{E}[(\bar{X} + \tilde{X})u_j] = \mathrm{E}[\bar{X}u_j]. \tag{4.148}$$

The last term in zero since $\tilde{X}$ is orthogonal to all basis vectors in the space by definition.

Now, back to the general DLM problem, where we have the random vector set $\tilde{\mathbf{Y}}_n$. If we say the length of each $\mathbf{Y}_i$ is m then we have nm random variables that make up the $\boldsymbol{\Upsilon}_n$ vector space and nm basis random variables that also make up the space. Let us denote these basis random variables as,

$$\{\{u_{11}, ..., u_{1m}\}, ..., \{u_{n1}, ..., u_{nm}\}\} = \{\mathbf{u}_1, ..., \mathbf{u}_n\}. \tag{4.149}$$

We know, from the argument made earlier in this section, that any component of the MMSE estimator $\mathbf{X}_n^n$ for Gaussian distributed $\mathbf{X}_i$ and $\mathbf{Y}_i$ is the linear combination of these basis vectors. And now we know, from our discussion of basis random variables and that $\mathbf{X}_n^n$ is the orthogonal projection of $\mathbf{X}_n$ onto the $\boldsymbol{\Upsilon}_n$ vector space, what the coefficients that multiply these basis random variables should be. For instance, for the jth component of $\mathbf{X}_n^n$, the coefficient that multiplies the basis random variable u_{ts} (any arbitrary t and s) is $\mathrm{E}[X_{nj}u_{ts}]$ where we define X_{nj} as the jth component of $\mathbf{X}_n$. Using some linear algebra (instead of writing a double sum) we may write this linear combination as,

$$\mathbf{X}_n^n = \sum_{j=1}^{n} \mathrm{E}[\mathbf{X}_n \mathbf{u}_j^T]\mathbf{u}_j. \tag{4.150}$$

This equation will now be utilized to finally justify the form of Equation 4.116. We may break it into two terms as follows:

$$\mathbf{X}_n^n = \sum_{j=1}^{n-1} \mathrm{E}[\mathbf{X}_n \mathbf{u}_j^T]\mathbf{u}_j + \mathrm{E}[\mathbf{X}_n \mathbf{u}_n^T]\mathbf{u}_n. \tag{4.151}$$

Then, inserting the evolution equation into the first term we have,

$$\mathbf{X}_n^n = \sum_{j=1}^{n-1} \mathrm{E}[(\mathbf{G}_n \mathbf{X}_{n-1} + \mathbf{w}_n)\mathbf{u}_j^T]\mathbf{u}_j + \mathrm{E}[\mathbf{X}_n \mathbf{u}_n^T]\mathbf{u}_n. \tag{4.152}$$

Since $\mathbf{w}_n$ is uncorrelated to $\mathbf{u}_j$ for $j < n$ we have,

$$\mathbf{X}_n^n = \mathbf{G}_n \sum_{j=1}^{n-1} \mathrm{E}[\mathbf{X}_{n-1} \mathbf{u}_j^T]\mathbf{u}_j + \mathrm{E}[\mathbf{X}_n \mathbf{u}_n^T]\mathbf{u}_n \tag{4.153}$$

$$= \mathbf{G}_n \mathbf{X}_{n-1}^{n-1} + \mathrm{E}[\mathbf{X}_n \mathbf{u}_n^T]\mathbf{u}_n \tag{4.154}$$

$$= \mathbf{X}_n^{n-1} + \mathrm{E}[\mathbf{X}_n \mathbf{u}_n^T]\mathbf{u}_n. \tag{4.155}$$

The second step is made noticing that $\mathbf{X}_{n-1}^{n-1}$ appears, as defined by Equation 4.150. However, we notice that this step can only be made if we assume that the basis random variables that make up the $\boldsymbol{\Upsilon}_n$ space can be selected such that they are all of the basis random variables of the $\boldsymbol{\Upsilon}_{n-1}$ space plus one additional basis random variable. This can indeed be done and is in fact how the Gram-Schmidt process works naturally, so it is actually not an assumption at all. We see now from Equation 4.155 that the first term $\mathbf{X}_n^{n-1}$ is the same as the first term of Equation 4.116 that we are trying to justify. So, what remains is to show that the second term can be given as some linear operation on $\mathbf{Y}_n - \mathbf{F}_n \mathbf{X}_n^{n-1}$.

We note that the elements of the second term have to be orthogonal to all the elements of the first term since the basis random variables are all orthogonal to each other. More specifically, its elements are orthogonal to the $\boldsymbol{\Upsilon}_{n-1}$ vector space. From this we can see that what is precisely needed as the second term is some linear operation on the components of the elements of $\mathbf{Y}_n$ which are orthogonal to the $\boldsymbol{\Upsilon}_{n-1}$ vector space. We define the linear operation as $\mathbf{K}'_n$ and the part of $\mathbf{Y}_n$ orthogonal to $\boldsymbol{\Upsilon}_{n-1}$ as $\tilde{\mathbf{Y}}_n$ (this $\tilde{\mathbf{Y}}_n$ should not be confused with the collection of random vectors we have also denoted as $\tilde{\mathbf{Y}}_n$ in this thesis). Putting this into Equation 4.155 we have,

$$\mathbf{X}_n^n = \mathbf{X}_n^{n-1} + \mathbf{K}'_n \tilde{\mathbf{Y}}_n. \tag{4.156}$$

With the trick of Equation 4.138, we can write explicitly $\mathbf{Y}_n$ as the sum of $\bar{\mathbf{Y}}_n$ and $\tilde{\mathbf{Y}}_n$, where the elements of $\bar{\mathbf{Y}}_n$ are the components of the elements of $\mathbf{Y}_n$ in the $\boldsymbol{\Upsilon}_{n-1}$ vector space and $\tilde{\mathbf{Y}}_n$ is as defined above. We have,

$$\mathbf{Y}_n = \bar{\mathbf{Y}}_n + \tilde{\mathbf{Y}}_n. \tag{4.157}$$

Now, since the elements of $\mathbf{X}_n^{n-1}$ are the components of the elements of $\mathbf{X}_n$ in $\boldsymbol{\Upsilon}_{n-1}$ space, $\mathbf{F}_n\mathbf{X}_n^{n-1}$ is $\bar{\mathbf{Y}}_n$. Therefore, $\tilde{\mathbf{Y}}_n$ is given as,

$$\tilde{\mathbf{Y}}_n = \mathbf{Y}_n - \mathbf{F}_n\mathbf{X}_n^{n-1}, \tag{4.158}$$

and so we have completely justified the form of Equation 4.116 since we have,

$$\mathbf{X}_n^n = \mathbf{X}_n^{n-1} + \mathbf{K}_n'(\mathbf{Y}_n - \mathbf{F}_n\mathbf{X}_n^{n-1}). \tag{4.159}$$

To wrap this up, if the argument made about how $\bar{\mathbf{Y}}_n = \mathbf{F}_n\mathbf{X}_n^{n-1}$ is unclear or unsatisfactory, we can explicitly verify that $\mathbf{Y}_n - \mathbf{F}_n\mathbf{X}_n^{n-1}$ is the orthogonal part of $\mathbf{Y}_n$ to the $\boldsymbol{\Upsilon}_{n-1}$ space by showing that each of its elements is orthogonal to every basis random variable in the $\boldsymbol{\Upsilon}_{n-1}$ space. Mathematically, showing that

$$\mathrm{E}[(\mathbf{Y}_n - \mathbf{F}_n\mathbf{X}_n^{n-1})\mathbf{u}_j^T] = 0, \tag{4.160}$$

for all $j = 1, ..., n-1$. If this is shown to be true then $\mathbf{F}_n\mathbf{X}_n^{n-1}$ is indeed $\bar{\mathbf{Y}}_n$. By inserting the observation equation into this we have,

$$\mathrm{E}[(\mathbf{Y}_n - \mathbf{F}_n\mathbf{X}_n^{n-1})\mathbf{u}_j^T] = \mathrm{E}[(\mathbf{F}_n\mathbf{X}_n + \mathbf{v}_n - \mathbf{F}_n\mathbf{X}_n^{n-1})\mathbf{u}_j^T] \tag{4.161}$$

$$= \mathbf{F}_n\mathrm{E}[(\mathbf{X}_n - \mathbf{X}_n^{n-1})\mathbf{u}_j^T]. \tag{4.162}$$

We notice that the elements of $\mathbf{X}_n - \mathbf{X}_n^{n-1}$ are the components of the elements of $\mathbf{X}_n$ orthogonal to the $\boldsymbol{\Upsilon}_{n-1}$ space and that therefore each of its elements are orthogonal to all basis random variables that make up the $\boldsymbol{\Upsilon}_{n-1}$ space. Therefore this expectation is zero and we have verified that $\mathbf{Y}_n - \mathbf{F}_n\mathbf{X}_n^{n-1}$ is indeed the part of $\mathbf{Y}_n$ orthogonal to the $\boldsymbol{\Upsilon}_{n-1}$ space.

4.5.5 Cost Function Theorem

In this section, we give the cost function for DLM estimation and prove that it results in the filtering, smoothing, and prediction recurrence relations. For the filtering part, we cite (Sorenson, 1970). We have the following cost function theorem: given data $\tilde{\mathbf{y}}_n$, the cost function for DLM estimation is given as a function of $\mathbf{x}_1, ..., \mathbf{x}_n, ..., \mathbf{x}_p$ for some $p \geq n$ as,

$$L_n(\mathbf{x}_1, ..., \mathbf{x}_n, ..., \mathbf{x}_p) = \frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_1^0)^T \mathbf{P}_1^{0^{-1}}(\mathbf{x}_1 - \mathbf{x}_1^0) + \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i)^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i) + \frac{1}{2}\sum_{i=2}^{p}(\mathbf{x}_i - \mathbf{G}_i\mathbf{x}_{i-1})^T \mathbf{W}_i^{-1}(\mathbf{x}_i - \mathbf{G}_i\mathbf{x}_{i-1}). \quad (4.163)$$

What is meant by this is that the values of $\mathbf{x}_1, ..., \mathbf{x}_n, ..., \mathbf{x}_p$ that minimize this function are the estimates that we have previously defined as $\mathbf{x}_1^n, \mathbf{x}_2^n ..., \mathbf{x}_{n-1}^n, \mathbf{x}_n^n, \mathbf{x}_{n+1}^n, ..., \mathbf{x}_{p-1}^n, \mathbf{x}_p^n$ which take into account all of the given data $\tilde{\mathbf{y}}_n$. The estimates $\mathbf{x}_1^n$ through $\mathbf{x}_{n-1}^n$ are smoothed estimates, the estimate $\mathbf{x}_n^n$ is a filtered estimate, and the estimates $\mathbf{x}_{n+1}^n$ through $\mathbf{x}_p^n$ are predictions. We note that this is a multivariate quadratic function with all positive quadratic terms, so it has a single extremum that is a minimum. The extremum $\mathbf{x}_1^n, \mathbf{x}_2^n ..., \mathbf{x}_{n-1}^n, \mathbf{x}_n^n, \mathbf{x}_{n+1}^n, ..., \mathbf{x}_{p-1}^n, \mathbf{x}_p^n$ can be found by solving the system of equations where each partial derivative is set to zero. However, this problem is too complicated to be solved naively like this, so instead we will take an approach that allows us to obtain recursive solutions, which we will see are consistent with the filtering, smoothing, and prediction recurrence relations for the DLM.

We also note that the cost function given here is effectively a least squares cost function. The first two terms are actually identical to the weighted RLS cost function given in Appendix L (a slight modification to the RLS cost function presented in Section 4.1.3). The only difference is that the second term here in Equation 4.163 is for multivariate data. The last term is then characterized by the general DLM evolution equation, which is of course not a characteristic of the MLR model.

4.5.5.1 Prediction Proof

First, we will get the proof that the cost function given in Equation 4.163 is consistent with the prediction recurrence relation out of the way since it is much simpler than the smoothing

and filtering cases. Assume that the extremum values of the cost function $\mathbf{x}_1^n, ..., \mathbf{x}_{p-1}^n$ are known. Plugging these into the cost function we have,

$$\ln(\mathbf{x}_1^n, ..., \mathbf{x}_{p-1}^n, \mathbf{x}_p) = constant + \frac{1}{2}(\mathbf{x}_p - \mathbf{G}_p\mathbf{x}_{p-1})^T\mathbf{W}_p^{-1}(\mathbf{x}_p - \mathbf{G}_p\mathbf{x}_{p-1}), \tag{4.164}$$

where $\mathbf{x}_p$ is the only unspecified variable. The extremum value $\mathbf{x}_p^n$ that minimizes this function is clearly

$$\mathbf{x}_p^n = \mathbf{G}_p\mathbf{x}_{p-1}^n, \tag{4.165}$$

the prediction estimate recurrence relation already seen in this thesis. To see that this works for all prediction estimates defined by the cost function, consider looking at the case $p = n+1$, then $p = n+2$, and so on. The prediction estimate recurrence relation would be found simply as shown here for all p greater than n.

4.5.5.2 Three Results

Throughout the next proofs carried out for the filtering and smoothing problems, three results are used. These results are given here in this section and proven in Appendix R. So, rather than solving these problems as we need them for the filtering and smoothing problems we will refer to them here.

Result 1

The first result is the minima of the function:

$$F(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{a})^T\mathbf{A}(\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{b} - \mathbf{B}\mathbf{x})^T\mathbf{C}(\mathbf{b} - \mathbf{B}\mathbf{x}), \tag{4.166}$$

where $\mathbf{a}$ and $\mathbf{b}$ are vectors, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are matrices, and $\mathbf{A}$ and $\mathbf{C}$ are restricted to be symmetric. The location of the minimum, which we define as $\mathbf{x}_m$, can be written as,

$$\mathbf{x}_m = \mathbf{a} + \mathbf{A}^{-1}\mathbf{B}^T(\mathbf{C}^{-1} + \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)^{-1}(\mathbf{b} - \mathbf{B}\mathbf{a}). \tag{4.167}$$

Result 2

The second result is the minima of the function

$$F(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{a})^T\mathbf{A}(\mathbf{x}-\mathbf{a}) + \frac{1}{2}(\mathbf{b}-\mathbf{B}\mathbf{x})^T\mathbf{C}(\mathbf{b}-\mathbf{B}\mathbf{x}) + \frac{1}{2}(\mathbf{d}-\mathbf{D}\mathbf{x})^T\mathbf{E}(\mathbf{d}-\mathbf{D}\mathbf{x}), \quad (4.168)$$

where $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{d}$ are vectors, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{E}$ are matrices, and $\mathbf{A}$, $\mathbf{C}$, and $\mathbf{E}$ are restricted to be symmetric. The location $\mathbf{x}_m$ of the minimum can be written in the following way, which is only a convenient way to write it for the problem we use it for in this section, as,

$$\mathbf{x}_m = \mathbf{Z}^{-1}\mathbf{Y} + \mathbf{Z}^{-1}\mathbf{D}^T(\mathbf{E}^{-1} + \mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T)^{-1}(\mathbf{d} - \mathbf{D}\mathbf{Z}^{-1}\mathbf{Y}), \quad (4.169)$$

where $\mathbf{Z} = \mathbf{A} + \mathbf{B}^T\mathbf{C}\mathbf{B}$ and $\mathbf{Y} = \mathbf{A}\mathbf{a} + \mathbf{B}^T\mathbf{C}\mathbf{b}$.

Result 3

For the third result consider the following function:

$$F(\mathbf{x}_1, \mathbf{x}_2) = k + \frac{1}{2}(\mathbf{x}_1 - \mathbf{a})\mathbf{A}(\mathbf{x}_1 - \mathbf{a}) + \frac{1}{2}(\mathbf{b} - \mathbf{B}\mathbf{x}_2)^T\mathbf{C}(\mathbf{b} - \mathbf{B}\mathbf{x}_2) + \frac{1}{2}(\mathbf{x}_2 - \mathbf{D}\mathbf{x}_1)^T\mathbf{E}(\mathbf{x}_2 - \mathbf{D}\mathbf{x}_1), \quad (4.170)$$

where k is a constant, $\mathbf{a}$ and $\mathbf{b}$ are vectors, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{E}$ are matrices, and $\mathbf{A}$ and $\mathbf{E}$ are restricted to be symmetric. Let us define the single extremum of this function to be $(\mathbf{x}_{1m}, \mathbf{x}_{2m})$. If we set the derivative of this function with respect to $\mathbf{x}_1$ equal to zero, solve for $\mathbf{x}_{1m}$ (as a function of $\mathbf{x}_2$), and then plug the result back into the original function for $\mathbf{x}_1$, then we have a function of only $\mathbf{x}_2$ left over, with an extremum at $\mathbf{x}_{2m}$. This is just basic calculus. So, the third result that we give is that the function $F(\mathbf{x}_{1m}(\mathbf{x}_2), \mathbf{x}_2)$ (where we define $\mathbf{x}_{1m}(\mathbf{x}_2)$ as the $\mathbf{x}_{1m}$ function of $\mathbf{x}_2$ found as just described) that results from this procedure is

$$F(\mathbf{x}_{1m}(\mathbf{x}_2), \mathbf{x}_2) = k + \frac{1}{2}(\mathbf{b} - \mathbf{B}\mathbf{x}_2)^T\mathbf{C}(\mathbf{b} - \mathbf{B}\mathbf{x}_2) + \frac{1}{2}(\mathbf{x}_2 - \mathbf{D}\mathbf{a})^T\mathbf{P}^{-1}(\mathbf{x}_2 - \mathbf{D}\mathbf{a}), \quad (4.171)$$

where $\mathbf{P} = \mathbf{E}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T$. Again, the proofs of these results are given in Appendix R.

4.5.5.3 Filtering Estimates Proof

In this section, we prove that the cost function given in Equation 4.163 is consistent with the filtering estimate recurrence relations. We start by writing out the case where $n = 1$ and $p = 1$ (i.e. the cost function $L_1(\mathbf{x}_1)$ that has the extremum $\mathbf{x}_1^1$):

$$L_1(\mathbf{x}_1) = \frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_1^0)^T \mathbf{P}_1^{0^{-1}}(\mathbf{x}_1 - \mathbf{x}_1^0) + \frac{1}{2}(\mathbf{y}_1 - \mathbf{F}_1\mathbf{x}_1)^T \mathbf{V}_1^{-1}(\mathbf{y}_1 - \mathbf{F}_1\mathbf{x}_1). \tag{4.172}$$

As an aside, it can be quickly verified that the filtering recurrence relation is obtained for this individual case by using the first result in Section 4.5.5.2. It is seen that the minimum of $L_1(\mathbf{x}_1)$ is at $\mathbf{x}_1^1 = \mathbf{x}_1^0 + \mathbf{P}_1^0\mathbf{F}_1^T(\mathbf{V}_1 + \mathbf{F}_1\mathbf{P}_1^0\mathbf{F}_1^T)^{-1}(\mathbf{y}_1 - \mathbf{F}_1\mathbf{x}_1^0)$, which is the correct filtering solution that we have already seen for $\mathbf{x}_1^1$.

Now, to begin to tackle the general case, we first perform a quadratic Taylor series expansion of $L_1(\mathbf{x}_1)$ about $\mathbf{x}_1^1$ (see Appendix S for details on the multivariate Taylor series expansion). We have,

$$\begin{aligned} L_1(\mathbf{x}_1) = {} & \frac{1}{2}(\mathbf{x}_1^1 - \mathbf{x}_1^0)^T \mathbf{P}_1^{0^{-1}}(\mathbf{x}_1^1 - \mathbf{x}_1^0) + \frac{1}{2}(\mathbf{y}_1 - \mathbf{F}_1\mathbf{x}_1^1)^T \mathbf{V}_1^{-1}(\mathbf{y}_1 - \mathbf{F}_1\mathbf{x}_1^1) \\ & + \frac{\partial L_1(\mathbf{x}_1)}{\partial \mathbf{x}_1}\Big|_{\mathbf{x}_1 = \mathbf{x}_1^1}(\mathbf{x}_1 - \mathbf{x}_1^1) + \frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_1^1)^T \frac{\partial}{\partial \mathbf{x}_1}\left(\frac{\partial L_1(\mathbf{x}_1)}{\partial \mathbf{x}_1}\right)^T\Big|_{\mathbf{x}_1 = \mathbf{x}_1^1}(\mathbf{x}_1 - \mathbf{x}_1^1). \end{aligned} \tag{4.173}$$

Let us denote the sum of the first two constant terms as C_1 like so:

$$C_1 = \frac{1}{2}(\mathbf{x}_1^1 - \mathbf{x}_1^0)^T \mathbf{P}_1^{0^{-1}}(\mathbf{x}_1^1 - \mathbf{x}_1^0) + \frac{1}{2}(\mathbf{y}_1 - \mathbf{F}_1\mathbf{x}_1^1)^T \mathbf{V}_1^{-1}(\mathbf{y}_1 - \mathbf{F}_1\mathbf{x}_1^1). \tag{4.174}$$

The derivatives are found to be,

$$\frac{\partial L_1(\mathbf{x}_1)}{\partial \mathbf{x}_1} = (\mathbf{x}_1 - \mathbf{x}_1^0)^T \mathbf{P}_1^{0^{-1}} - (\mathbf{y}_1 - \mathbf{F}_1\mathbf{x}_1)^T \mathbf{V}_1^{-1}\mathbf{F}_1, \tag{4.175}$$

and

$$\frac{\partial}{\partial \mathbf{x}_1}\left(\frac{\partial L_1(\mathbf{x}_1)}{\partial \mathbf{x}_1}\right)^T = \mathbf{P}_1^{0^{-1}} - \mathbf{F}_1^T \mathbf{V}_1^{-1}\mathbf{F}_1. \tag{4.176}$$

We note that $\frac{\partial L_1(\mathbf{x}_1)}{\partial \mathbf{x}_1}\Big|_{\mathbf{x}_1=\mathbf{x}_1^1}$ is clearly zero since the point $\mathbf{x}_1^1$ is the location of the minimum of this function by definition. So, the new expression of $L_1(\mathbf{x}_1)$ obtained from this quadratic Taylor series expansion is

$$L_1(\mathbf{x}_1) = C_1 + \frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_1^1)^T(\mathbf{P}_1^{0-1} - \mathbf{F}_1^T\mathbf{V}_1^{-1}\mathbf{F}_1)(\mathbf{x}_1 - \mathbf{x}_1^1). \tag{4.177}$$

This is not an approximation since Equation 4.173 is only quadratic in $\mathbf{x}_1$ and clearly third order derivatives are zero. Now, we know from the DLM theory that $\mathbf{P}_i^i = \mathbf{P}_i^{i-1} - \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T + \mathbf{V}_i)^{-1}\mathbf{F}_i\mathbf{P}_i^{i-1} = (\mathbf{P}_i^{i-1^{-1}} - \mathbf{F}_i^T\mathbf{V}_i^{-1}\mathbf{F}_i)^{-1}$. So, we will simplify this equation slightly by defining $\mathbf{P}_1^{0-1} - \mathbf{F}_1^T\mathbf{V}_1^{-1}\mathbf{F}_1$ as $\mathbf{P}_1^{1-1}$. We have,

$$L_1(\mathbf{x}_1) = C_1 + \frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_1^1)^T\mathbf{P}_1^{1-1}(\mathbf{x}_1 - \mathbf{x}_1^1). \tag{4.178}$$

The next thing we will do is write $L_2(\mathbf{x}_1, \mathbf{x}_2)$ (this is the $n = p = 2$ case now). We see from the definition of the cost function that we can write $L_2(\mathbf{x}_1, \mathbf{x}_2)$ as the sum of $L_1(\mathbf{x}_1)$ and two additional terms. So we have,

$$\begin{aligned} L_2(\mathbf{x}_1, \mathbf{x}_2) = C_1 &+ \frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_1^1)^T\mathbf{P}_1^{1-1}(\mathbf{x}_1 - \mathbf{x}_1^1) + \frac{1}{2}(\mathbf{y}_2 - \mathbf{F}_2\mathbf{x}_2)^T\mathbf{V}_2^{-1}(\mathbf{y}_2 - \mathbf{F}_2\mathbf{x}_2) \\ &+ \frac{1}{2}(\mathbf{x}_2 - \mathbf{G}_2\mathbf{x}_1)^T\mathbf{W}_2^{-1}(\mathbf{x}_2 - \mathbf{G}_2\mathbf{x}_1), \end{aligned} \tag{4.179}$$

where the last two terms are the two additional terms we are referring to. Now, using the third result of Section 4.5.5.2 we may write the function $L_2(\mathbf{x}_1^2(\mathbf{x}_2), \mathbf{x}_2)$ as,

$$L_2(\mathbf{x}_1^2(\mathbf{x}_2), \mathbf{x}_2) = C_1 + \frac{1}{2}(\mathbf{y}_2 - \mathbf{F}_2\mathbf{x}_2)^T\mathbf{V}_2^{-1}(\mathbf{y}_2 - \mathbf{F}_2\mathbf{x}_2) + \frac{1}{2}(\mathbf{x}_2 - \mathbf{G}_2\mathbf{x}_1^1)^T\mathbf{A}^{-1}(\mathbf{x}_2 - \mathbf{G}_2\mathbf{x}_1^1), \tag{4.180}$$

where $\mathbf{A} = \mathbf{W}_2 + \mathbf{G}_2\mathbf{P}_1^1\mathbf{G}_2^T$. Recall from the thrid result of Section 4.5.5.2 that this means that the extremum of this function is also equal to $\mathbf{x}_2^2$. To simplify this slightly, we have $\mathbf{G}_2\mathbf{x}_1^1 = \mathbf{x}_2^1$ by recalling the prediction estimate, and by carrying out the evaluation of $\mathbf{P}_2^1 = \text{Cov}[\mathbf{X}_2^1 - \mathbf{X}_2]$ (which has been done in Section 4.5.3.3 for the general $\mathbf{P}_i^{i-1}$ case) we notice that $\mathbf{A} = \mathbf{P}_2^1$. So, $L_2(\mathbf{x}_1^2(\mathbf{x}_2), \mathbf{x}_2)$ can be simplified to

$$L_2(\mathbf{x}_1^2(\mathbf{x}_2), \mathbf{x}_2) = C_1 + \frac{1}{2}(\mathbf{y}_2 - \mathbf{F}_2\mathbf{x}_2)^T\mathbf{V}_2^{-1}(\mathbf{y}_2 - \mathbf{F}_2\mathbf{x}_2) + \frac{1}{2}(\mathbf{x}_2 - \mathbf{x}_2^1)^T\mathbf{P}_2^{1-1}(\mathbf{x}_2 - \mathbf{x}_2^1). \tag{4.181}$$

Now, what we notice here is that this function of $\mathbf{x}_2$ is essentially identical to the form of $L_1(\mathbf{x}_1)$ in Equation 4.172. The difference is that all the indices are incremented by 1. Recall that the filtering recurrence relation for $\mathbf{x}_1^1$ was found with Equation 4.172, so because the equation here is essentially identical we can see that the extremum $\mathbf{x}_2^2$ of this function is also the same as the filtering recurrence relation for $\mathbf{x}_2^2$.

What we will do now is essentially repeat the procedure we have followed so far to generalize this argument to any case $n = p = i$. For instance, for the next case of $n = p = 3$ we would follow the same procedure of Taylor series expanding $L_2(\mathbf{x}_1^2(\mathbf{x}_2), \mathbf{x}_2)$ about $\mathbf{x}_2^2$, add the two additional terms to obtain $L_3(\mathbf{x}_1^2(\mathbf{x}_2), \mathbf{x}_2, \mathbf{x}_3)$, and then use the third result of Section 4.5.5.2 to obtain the function of only $\mathbf{x}_3$ where the extremum value of $\mathbf{x}_2$ (i.e. $\mathbf{x}_2^3$) as a function of $\mathbf{x}_3$ has been solved for. Using our notation, we will have to denote this as $L_3(\mathbf{x}_1^2(\mathbf{x}_2^3(\mathbf{x}_3)), \mathbf{x}_2^3(\mathbf{x}_3), \mathbf{x}_3)$. We know because of the identical procedure just carried out for $n = p = 2$ that this will result in:

$$L_3(\mathbf{x}_1^2(\mathbf{x}_2^3(\mathbf{x}_3)), \mathbf{x}_2^3(\mathbf{x}_3), \mathbf{x}_3) = C_1 + C_2 + \frac{1}{2}(\mathbf{y}_3 - \mathbf{F}_3\mathbf{x}_3)^T\mathbf{V}_3^{-1}(\mathbf{y}_3 - \mathbf{F}_3\mathbf{x}_3) + \frac{1}{2}(\mathbf{x}_3 - \mathbf{x}_3^2)^T\mathbf{P}_3^{2-1}(\mathbf{x}_3 - \mathbf{x}_3^2), \tag{4.182}$$

where C_2 is another constant. This function is again in identical form to $L_1(\mathbf{x}_1)$ in Equation 4.172 and $L_2(\mathbf{x}_1^2(\mathbf{x}_2), \mathbf{x}_2)$ in Equation 4.181 with only the indices incremented. So, we again find that the extremum $\mathbf{x}_3^3$ of this function is the same as the filtering recurrence relation for $\mathbf{x}_3^3$. But, unfortunately we have not been completely rigorous with this extension to the $n = p = 3$ case yet. We have to verify that the minimum of this function $L_3(\mathbf{x}_1^2(\mathbf{x}_2^3(\mathbf{x}_3)), \mathbf{x}_2^3(\mathbf{x}_3), \mathbf{x}_3)$ is the same as the minimum of the function $L_3(\mathbf{x}_1^3(\mathbf{x}_3), \mathbf{x}_2^3(\mathbf{x}_3), \mathbf{x}_3)$ (whose minimum we know to be $\mathbf{x}_3^3$) where the true extremum values $\mathbf{x}_1^3$ and $\mathbf{x}_2^3$ as a function of $\mathbf{x}_3$ have been plugged in. To be clear, these functions $\mathbf{x}_1^3(\mathbf{x}_3)$ and $\mathbf{x}_2^3(\mathbf{x}_3)$ would be found by solving the following simultaneous equations:

$$\frac{\partial L_3(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)}{\partial \mathbf{x}_1} = 0 \text{ and } \frac{\partial L_3(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)}{\partial \mathbf{x}_2} = 0. \tag{4.183}$$

Now, we can verify this it true by considering going about solving $\frac{\partial L_3(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)}{\partial \mathbf{x}_1} = 0$ for $\mathbf{x}_1$. The result, in general, would be $\mathbf{x}_1^3$ as function of $\mathbf{x}_2$ and $\mathbf{x}_3$. But, for our specific cost function

$L_3(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ we see pretty clearly that it would only be a function of $\mathbf{x}_2$, and furthermore that it would be the same function of $\mathbf{x}_2$ as the function $\mathbf{x}_1^2(\mathbf{x}_2)$ that we have seen in this section, which was obtained from:

$$\frac{\partial L_2(\mathbf{x}_1, \mathbf{x}_2)}{\partial \mathbf{x}_1} = 0. \tag{4.184}$$

To express this mathematically, if we define this function for $\mathbf{x}_1^3$ resulting from $\frac{\partial L_3(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)}{\partial \mathbf{x}_1} = 0$ as $\mathbf{x}_1^3(\mathbf{x}_2)$, then we can say that we have $\mathbf{x}_1^3(\mathbf{x}_2) = \mathbf{x}_1^2(\mathbf{x}_2)$. To obtain what we have defined as $\mathbf{x}_1^3(\mathbf{x}_3)$ now, we simply plug in $\mathbf{x}_2^3(\mathbf{x}_3)$ into $\mathbf{x}_1^3(\mathbf{x}_2)$. So, we have $\mathbf{x}_1^3(\mathbf{x}_3) = \mathbf{x}_1^3(\mathbf{x}_2^3(\mathbf{x}_3)) = \mathbf{x}_1^2(\mathbf{x}_2^3(\mathbf{x}_3))$. Therefore, we have the result:

$$L_3(\mathbf{x}_1^2(\mathbf{x}_2^3(\mathbf{x}_3)), \mathbf{x}_2^3(\mathbf{x}_3), \mathbf{x}_3) = L_3(\mathbf{x}_1^3(\mathbf{x}_3), \mathbf{x}_2^3(\mathbf{x}_3), \mathbf{x}_3), \tag{4.185}$$

showing that these functions are the same. So finally, the extremum of them both is indeed $\mathbf{x}_3^3$.

For all the next cases $n = p = i$ for any $i > 3$, we are able to verify a similar result by considering what the solutions to the simultaneous equation look like, as we have done above. So, we can see that the generalization of this procedure, of Taylor series expanding, adding the two terms, and using the third result in Section 4.5.5.2, to the case of $n = p = i$ for any i results in:

$$\begin{aligned} &L_i(\mathbf{x}_1^i(\mathbf{x}_i), \mathbf{x}_2^i(\mathbf{x}_i), ..., \mathbf{x}_{i-1}^i(\mathbf{x}_i), \mathbf{x}_i) = \\ &\qquad \sum_{j=1}^{i-1} C_j + \frac{1}{2}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i)^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i) + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_i^{i-1})^T\mathbf{P}_i^{i-1^{-1}}(\mathbf{x}_i - \mathbf{x}_i^{i-1}), \end{aligned} \tag{4.186}$$

where all C_j are constants. It is not really important what the constants are since this is a cost function, but in case this becomes useful to someone it is easy to see that they are given as,

$$C_j = \frac{1}{2}(\mathbf{x}_j^j - \mathbf{x}_j^{j-1})^T\mathbf{P}_j^{j-1^{-1}}(\mathbf{x}_j^j - \mathbf{x}_j^{j-1}) + \frac{1}{2}(\mathbf{y}_j - \mathbf{F}_j\mathbf{x}_j^j)^T\mathbf{V}_j^{-1}(\mathbf{y}_j - \mathbf{F}_j\mathbf{x}_j^j). \tag{4.187}$$

Now, as we have similarly said several times for functions in this form, we know from using the first result of Section 4.5.5.2 that the minimum point of the function in Equation 4.186 is the filtering recurrence relation for $\mathbf{x}_i^i$. It is found to be,

$$\mathbf{x}_i^i = \mathbf{x}_i^{i-1} + \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{V}_i + \mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T)^{-1}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i^{i-1}), \tag{4.188}$$

which is indeed the filtering recurrence relation that has been shown and derived previously.

4.5.5.4 Smoothing Estimates Proof

In this section, we prove that the cost function given by Equation 4.163 is consistent with the smoothing recurrence relations for the DLM. Consider writing the following:

$$L_n(\mathbf{x}_1^n(\mathbf{x}_i), \mathbf{x}_2^n(\mathbf{x}_i), ..., \mathbf{x}_{i-1}^n(\mathbf{x}_i), \mathbf{x}_i, \mathbf{x}_{i+1}, ..., \mathbf{x}_n). \tag{4.189}$$

That is, the cost function $L_n(\mathbf{x}_1, ..., \mathbf{x}_n)$ where the functions of

$$\mathbf{x}_1^n(\mathbf{x}_i), \mathbf{x}_2^n(\mathbf{x}_i), ..., \text{ and } \mathbf{x}_{i-1}^n(\mathbf{x}_i), \tag{4.190}$$

that have been obtained by solving the simultaneous equations

$$\frac{\partial F_n(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)}{\partial \mathbf{x}_1} = 0, \frac{\partial F_n(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)}{\partial \mathbf{x}_2} = 0, ..., \text{ and } \frac{\partial F_n(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)}{\partial \mathbf{x}_{i-1}} = 0, \tag{4.191}$$

are plugged in for $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{i-1}$. It is clear that the extremum point of this function is $(\mathbf{x}_i^n, \mathbf{x}_{i+1}^n, ..., \mathbf{x}_n^n)$.

Now, we notice that the functions of $\mathbf{x}_1^i(\mathbf{x}_i), \mathbf{x}_2^i(\mathbf{x}_i), ..., \mathbf{x}_{i-1}^i(\mathbf{x}_i)$, that are obtained from solving the simultaneous equations

$$\frac{\partial F_i(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i)}{\partial \mathbf{x}_1} = 0, \frac{\partial F_i(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i)}{\partial \mathbf{x}_2} = 0, ..., \text{ and } \frac{\partial F_i(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i)}{\partial \mathbf{x}_{i-1}} = 0, \tag{4.192}$$

would be the same as the functions $\mathbf{x}_1^n(\mathbf{x}_i), \mathbf{x}_2^n(\mathbf{x}_i), ..., \mathbf{x}_{i-1}^n(\mathbf{x}_i)$. To be clear, we can write,

$$\mathbf{x}_1^i(\mathbf{x}_i) = \mathbf{x}_1^n(\mathbf{x}_i), \mathbf{x}_2^i(\mathbf{x}_i) = \mathbf{x}_2^n(\mathbf{x}_i), ..., \text{ and } \mathbf{x}_{i-1}^i(\mathbf{x}_i) = \mathbf{x}_{i-1}^n(\mathbf{x}_i). \tag{4.193}$$

This is just due to the fact that the additional terms added for $L_n(\mathbf{x}_1, ..., \mathbf{x}_n)$ from $L_n(\mathbf{x}_1, ..., \mathbf{x}_i)$ do not depend on any of $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{i-1}$, so the simultaneous equations are the same in both cases. So, we see that we have,

$$L_n(\mathbf{x}_1^n(\mathbf{x}_i), \mathbf{x}_2^n(\mathbf{x}_i), ..., \mathbf{x}_{i-1}^n(\mathbf{x}_i), \mathbf{x}_i, \mathbf{x}_{i+1}, ..., \mathbf{x}_n) = L_n(\mathbf{x}_1^i(\mathbf{x}_i), \mathbf{x}_2^i(\mathbf{x}_i), ..., \mathbf{x}_{i-1}^i(\mathbf{x}_i), \mathbf{x}_i, \mathbf{x}_{i+1}, ..., \mathbf{x}_n). \tag{4.194}$$

This allows us to notice that we can write this $L_n(\mathbf{x}_1^n(\mathbf{x}_i), \mathbf{x}_2^n(\mathbf{x}_i), ..., \mathbf{x}_{i-1}^n(\mathbf{x}_i), \mathbf{x}_i, \mathbf{x}_{i+1}, ..., \mathbf{x}_n)$ function as the sum of $L_i(\mathbf{x}_1^i(\mathbf{x}_i), \mathbf{x}_2^i(\mathbf{x}_i), ..., \mathbf{x}_{i-1}^i(\mathbf{x}_i), \mathbf{x}_i)$ (which was specified in the last section) and two additional terms for each index $i+1$ through n. So, we have,

$$\begin{aligned} L_n(\mathbf{x}_1^i(\mathbf{x}_i),&\mathbf{x}_2^i(\mathbf{x}_i), ..., \mathbf{x}_{i-1}^i(\mathbf{x}_i), \mathbf{x}_i, \mathbf{x}_{i+1}, ..., \mathbf{x}_n) = \\ &\sum_{j=1}^{i-1} C_j + \frac{1}{2}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i)^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i) + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_i^{i-1})^T\mathbf{P}_i^{i-1^{-1}}(\mathbf{x}_i - \mathbf{x}_i^{i-1}) \\ &+ \frac{1}{2}(\mathbf{x}_{i+1} - \mathbf{G}_{i+1}\mathbf{x}_i)^T\mathbf{W}_{i+1}^{-1}(\mathbf{x}_{i+1} - \mathbf{G}_{i+1}\mathbf{x}_i) + \ldots, \end{aligned}$$

where the additional terms that are not functions of $\mathbf{x}_i$ have not been written (so we have only actually written one additional term). We now seek to find the point $(\mathbf{x}_i^n, \mathbf{x}_{i+1}^n, ..., \mathbf{x}_n^n)$ that minimizes this function. If we assume that $\mathbf{x}_{i+1}^n, ..., \mathbf{x}_n^n$ are known and only $\mathbf{x}_i^n$ is unknown then we can plug in $\mathbf{x}_{i+1}^n, ..., \mathbf{x}_n^n$ for $\mathbf{x}_{i+1}, ..., \mathbf{x}_n$ respectively and write the following:

$$\begin{aligned} L_n(\mathbf{x}_1^i(\mathbf{x}_i),&\mathbf{x}_2^i(\mathbf{x}_i), ..., \mathbf{x}_{i-1}^i(\mathbf{x}_i), \mathbf{x}_i, \mathbf{x}_{i+1}^n, ..., \mathbf{x}_n^n) = \\ &\frac{1}{2}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i)^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i) + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_i^{i-1})^T\mathbf{P}_i^{i-1^{-1}}(\mathbf{x}_i - \mathbf{x}_i^{i-1}) \\ &+ \frac{1}{2}(\mathbf{x}_{i+1}^n - \mathbf{G}_{i+1}\mathbf{x}_i)^T\mathbf{W}_{i+1}^{-1}(\mathbf{x}_{i+1}^n - \mathbf{G}_{i+1}\mathbf{x}_i) + \textit{constant}. \end{aligned}$$

Now we have a function of only $\mathbf{x}_i$. Using the second result of Section 4.5.5.2 we find the minimum point of this function to be,

$$\mathbf{x}_i^n = \mathbf{Z}^{-1}\mathbf{Y} + \mathbf{Z}^{-1}\mathbf{G}_{i+1}^T(\mathbf{W}_{i+1} + \mathbf{G}_{i+1}\mathbf{Z}^{-1}\mathbf{G}_{i+1}^T)^{-1}(\mathbf{x}_{i+1}^n - \mathbf{G}_{i+1}\mathbf{Z}^{-1}\mathbf{Y}), \tag{4.195}$$

where $\mathbf{Z} = \mathbf{P}_i^{i-1^{-1}} + \mathbf{F}_i^T \mathbf{V}_i^{-1} \mathbf{F}_i^T$ and $\mathbf{Y} = \mathbf{P}_i^{i-1^{-1}} \mathbf{x}_i^{i-1} + \mathbf{F}_i^T \mathbf{V}_i^{-1} \mathbf{y}_i$. We notice that $\mathbf{Z}^{-1}$ is equal to $\mathbf{P}_i^i$, $\mathbf{Z}^{-1}\mathbf{Y}$ is equal to $\mathbf{x}_i^i$, and $\mathbf{W}_{i+1} + \mathbf{G}_{i+1}\mathbf{Z}^{-1}\mathbf{G}_{i+1}^T$ is equal to $\mathbf{P}_{i+1}^i$. Therefore we have,

$$\mathbf{x}_i^n = \mathbf{x}_i^i + \mathbf{P}_i^i \mathbf{G}_{i+1}^T {\mathbf{P}_{i+1}^i}^{-1} (\mathbf{x}_{i+1}^n - \mathbf{x}_{i+1}^i), \tag{4.196}$$

which is the same smoothed estimate that has been shown and derived in previous sections. So, in this DLM cost function section we have given a least squares estimation cost function and proven that it gives the prediction, filtering, and smoothing recurrence relations for DLM estimation.

4.5.6 Maximum Likelihood Estimate

In this section, we show that the same DLM filtering and smoothing recurrence relations can be obtained by an estimate known as the marginal maximum likelihood estimate (MLE). We follow very closely to the seminal work of (Rauch et al., 1965) in this section, and we note that this work was how the solution to the smoothing problem was first obtained (i.e. the smoothing recurrence relations). The marginal MLE is obtained from maximizing the marginal distribution $p(\mathbf{x}_i|\tilde{\mathbf{y}}_n)$. The maximum of this, which we will denote as $\mathbf{x}_i^n$, is obtained as a solution to the equation of its derivative with respect to $\mathbf{x}_i$ equal to 0. This is opposed to the joint MLE where the maximum of the joint distribution $p(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i|\tilde{\mathbf{y}}_n)$ is found by solving i simultaneous equations of derivatives equal to zeros. According to (Rauch et al., 1965), this also results in the same filtering and smoothing recurrence relations we have seen. But, only the marginal MLE proof is shown in this section.

4.5.6.1 The Filtering Estimation Derivation

For the filtering problem, we have $n = i$. So, we want to find the maximum of the marginal distribution $p(\mathbf{x}_i|\tilde{\mathbf{y}}_i)$. We have already written $p(\mathbf{x}_i|\tilde{\mathbf{y}}_i)$ in a certain way with Bayes theorem in Section 4.5.2.4. Here we will find the same formula in a slightly different way, following what is shown in (Rauch et al., 1965) by continually applying the following property of joint distributions:

$$p(a,b) = p(a|b)p(b). \tag{4.197}$$

To start, we have,

$$p(\mathbf{x}_i|\tilde{\mathbf{y}}_i) = \frac{p(\mathbf{x}_i, \tilde{\mathbf{y}}_i)}{p(\tilde{\mathbf{y}}_i)}, \tag{4.198}$$

where $p(\tilde{\mathbf{y}}_i)$ is a constant. Then, for $p(\mathbf{x}_i, \tilde{\mathbf{y}}_i)$ we have,

$$p(\mathbf{x}_i, \tilde{\mathbf{y}}_i) = p(\mathbf{x}_i, \mathbf{y}_i, \tilde{\mathbf{y}}_{i-1}) = p(\mathbf{y}_i|\mathbf{x}_i, \tilde{\mathbf{y}}_{i-1})p(\mathbf{x}_i, \tilde{\mathbf{y}}_{i-1}), \tag{4.199}$$

where $p(\mathbf{y}_i|\mathbf{x}_i, \tilde{\mathbf{y}}_{i-1})$ is the same as the likelihood function in Section 4.5.2.4, and reduces to $p(\mathbf{y}_i|\mathbf{x}_i)$ as discussed in the section. For $p(\mathbf{x}_i, \tilde{\mathbf{y}}_{i-1})$ we have,

$$p(\mathbf{x}_i, \tilde{\mathbf{y}}_{i-1}) = p(\mathbf{x}_i|\tilde{\mathbf{y}}_{i-1})p(\tilde{\mathbf{y}}_{i-1}), \tag{4.200}$$

where $p(\tilde{\mathbf{y}}_{i-1})$ is a constant. So, putting all this together we have,

$$p(\mathbf{x}_i|\tilde{\mathbf{y}}_i) \propto p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{x}_i|\tilde{\mathbf{y}}_{i-1}), \tag{4.201}$$

which is the same as what we had in Section 4.5.2.4. Recall from this section that $p(\mathbf{y}_i|\mathbf{x}_i)$ (we have not used our likelihood notation of $\mathcal{L}(\mathbf{y}_i|\mathbf{x}_i)$ here because it is not convenient to do so) is written with a mean of $\mathbf{F}_i\mathbf{X}_i$ and covariance of $\mathbf{V}_i$, and that $p(\mathbf{x}_i|\tilde{\mathbf{y}}_{i-1})$ has mean $\mathbf{x}_i^{i-1}$ and covariance $\mathbf{P}_i^{i-1}$. To find the marginal MLE, we find the maximum of the logarithm of $p(\mathbf{x}_i|\tilde{\mathbf{y}}_i)$, which is given as,

$$\begin{aligned}\ln(p(\mathbf{x}_i|\tilde{\mathbf{y}}_i)) &= \ln(p(\mathbf{y}_i|\mathbf{x}_i)) + \ln(p(\mathbf{x}_i|\tilde{\mathbf{y}}_{i-1})) + \ln(constant) \\ &= -\frac{1}{2}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i)^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i) - \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_i^{i-1})^T\mathbf{P}_i^{i-1-1}(\mathbf{x}_i - \mathbf{x}_i^{i-1}) \\ &\quad + constant.\end{aligned} \tag{4.202}$$

The extremum of the negative of this function was already found in Section 4.5.5.3 (this was the section where we proved the cost function theorem for the filtering problem). Using this result we have the maximum of this function, or the marginal MLE given as,

$$\mathbf{x}_i^i = \mathbf{x}_i^{i-1} + \mathbf{P}_i^{i-1}\mathbf{F}_i^T(\mathbf{V}_i + \mathbf{F}_i\mathbf{P}_i^{i-1}\mathbf{F}_i^T)^{-1}(\mathbf{y}_i - \mathbf{F}_i\mathbf{x}_i^{i-1}), \tag{4.203}$$

which is the filtering recurrence relation we have now seen numerous times.

If we were going to look at this marginal MLE as a stand-alone derivation for the filtering recurrence relations, we would now also want to find the error covariance matrix $\mathbf{P}_i^i$ by evaluating $\text{Cov}[\mathbf{X}_i^i - \mathbf{X}_i]$ where $\mathbf{X}_i^i$ is the estimator of the estimate $\mathbf{x}_i^i$ obtained here. This has technically been done already in Section 4.5.3.3, so it is not repeated here.

4.5.6.2 The Smoothing Estimation Derivation

For the smoothing problem, we want to find the maximum $\mathbf{x}_i^n$ of the marginal distribution $p(\mathbf{x}_i|\tilde{\mathbf{y}}_n)$ where $i < n$. But, what we do instead is find the maximum of the joint probability distribution $p(\mathbf{x}_i, \mathbf{x}_{i+1}|\tilde{\mathbf{y}}_n)$, where the values that maximize this function are $\mathbf{x}_i^n$ and $\mathbf{x}_{i+1}^n$. This gives us another way to find $\mathbf{x}_i^n$.

Similarly to the last section, we follow what is done in (Rauch et al., 1965) and obtain an expression for this distribution by continually applying the $p(a, b) = p(a|b)p(b)$ property of joint distributions. To start, we have,

$$p(\mathbf{x}_i, \mathbf{x}_{i+1}|\tilde{\mathbf{y}}_n) = \frac{p(\mathbf{x}_i, \mathbf{x}_{i+1}, \tilde{\mathbf{y}}_n)}{p(\tilde{\mathbf{y}}_n)}, \tag{4.204}$$

where $p(\tilde{\mathbf{y}}_n)$ is a constant. For $p(\mathbf{x}_i, \mathbf{x}_{i+1}, \tilde{\mathbf{y}}_n)$ we have,

$$p(\mathbf{x}_i, \mathbf{x}_{i+1}, \tilde{\mathbf{y}}_n) = p(\mathbf{x}_i, \mathbf{x}_{i+1}, \mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n|\tilde{\mathbf{y}}_i)p(\tilde{\mathbf{y}}_i), \tag{4.205}$$

where $p(\tilde{\mathbf{y}}_i)$ is a constant. For $p(\mathbf{x}_i, \mathbf{x}_{i+1}, \mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n|\tilde{\mathbf{y}}_i)$ we have,

$$p(\mathbf{x}_i, \mathbf{x}_{i+1}, \mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n|\tilde{\mathbf{y}}_i) = p(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n|\mathbf{x}_i, \tilde{\mathbf{y}}_i)p(\mathbf{x}_i|\tilde{\mathbf{y}}_i), \tag{4.206}$$

where $p(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n|\mathbf{x}_i, \tilde{\mathbf{y}}_i)$ reduces to $p(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n|\mathbf{x}_i)$ because $\mathbf{x}_{i+1}$, $\mathbf{y}_{i+1}$, $\mathbf{y}_{i+2}$, ..., $\mathbf{y}_n$ given $\mathbf{x}_i$ is independent of $\tilde{\mathbf{y}}_i$ (we see this from the evolution and observation equations). So, for $p(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n|\mathbf{x}_i, \tilde{\mathbf{y}}_i)$ we have,

$$p(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n | \mathbf{x}_i, \tilde{\mathbf{y}}_i) = p(\mathbf{x}_{i+1}, \mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n | \mathbf{x}_i) \tag{4.207}$$

$$= p(\mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n | \mathbf{x}_{i+1}, \mathbf{x}_i) p(\mathbf{x}_{i+1} | \mathbf{x}_i), \tag{4.208}$$

where for $p(\mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n | \mathbf{x}_{i+1}, \mathbf{x}_i)$ we see from the DLM observation equation that we have,

$$p(\mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n | \mathbf{x}_{i+1}, \mathbf{x}_i) = p(\mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n | \mathbf{x}_{i+1}). \tag{4.209}$$

Putting this all together we finally have $p(\mathbf{x}_i, \mathbf{x}_{i+1} | \tilde{\mathbf{y}}_n)$ in the form we want given as,

$$p(\mathbf{x}_i, \mathbf{x}_{i+1} | \tilde{\mathbf{y}}_n) \propto p(\mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n | \mathbf{x}_{i+1}) p(\mathbf{x}_{i+1} | \mathbf{x}_i) p(\mathbf{x}_i | \tilde{\mathbf{y}}_i). \tag{4.210}$$

The $p(\mathbf{x}_{i+1} | \mathbf{x}_i)$ distribution has a mean of $\mathbf{G}_{i+1}\mathbf{x}_i$ and covariance of $\mathbf{W}_{i+1}$ and the $p(\mathbf{x}_i | \tilde{\mathbf{y}}_i)$ distribution is our filtering distribution with a mean of $\mathbf{x}_i^i$ and covariance of $\mathbf{P}_i^i$. We will assume that we know the value $\mathbf{x}_{i+1}^n$ (similarly to how we assumed the higher index smoothing estimates were known in Section 4.5.5.4) so that the only job is to find the other value $\mathbf{x}_i^n$ that makes up the extremum point of $p(\mathbf{x}_i, \mathbf{x}_{i+1} | \tilde{\mathbf{y}}_n)$. With this assumption we see when we write it out that we can ignore the $p(\mathbf{y}_{i+1}, \mathbf{y}_{i+2}, ..., \mathbf{y}_n | \mathbf{x}_{i+1})$ distribution when finding the extremum value $\mathbf{x}_i^n$ because it does not depend of $\mathbf{x}_i$. So, to find $\mathbf{x}_i^n$, we find the maximum of the logarithm of $p(\mathbf{x}_i, \mathbf{x}_{i+1} | \tilde{\mathbf{y}}_n)$ with $\mathbf{x}_{i+1}^n$ plugged into the function for $\mathbf{x}_{i+1}$. The logarithm is given as,

$$\ln(p(\mathbf{x}_i, \mathbf{x}_{i+1}^n | \tilde{\mathbf{y}}_n)) = \ln(p(\mathbf{x}_{i+1}^n | \mathbf{x}_i)) + \ln(p(\mathbf{x}_i | \tilde{\mathbf{y}}_i)) + \ln(constant) \tag{4.211}$$

$$= -\frac{1}{2}(\mathbf{x}_{i+1}^n - \mathbf{G}_{i+1}\mathbf{x}_i)^T \mathbf{W}_{i+1}^{-1} (\mathbf{x}_{i+1}^n - \mathbf{G}_{i+1}\mathbf{x}_i) \tag{4.212}$$

$$- \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_i^i)^T \mathbf{P}_i^{i-1} (\mathbf{x}_i - \mathbf{x}_i^i) + constant. \tag{4.213}$$

Using the first result of Section 4.5.5.2, we find the maximum of this function to be,

$$\mathbf{x}_i^n = \mathbf{x}_i^i + \mathbf{P}_i^i \mathbf{G}_{i+1}^T {\mathbf{P}_{i+1}^i}^{-1} (\mathbf{x}_{i+1}^n - \mathbf{x}_{i+1}^i), \tag{4.214}$$

where $\mathbf{P}_{i+1}^{i} = \mathbf{W}_{i+1} + \mathbf{G}_{i+1}^{T}\mathbf{P}_{i}^{i}\mathbf{G}_{i+1}$ and $\mathbf{x}_{i+1}^{i} = \mathbf{G}_{i+1}\mathbf{x}_{i}^{i}$. This is indeed the smoothing recurrence relation we have seen many times now.

Likewise to the end of the last section on the filtering estimate, in order for this MLE derivation to be stand-alone for the smoothing recurrence relations, we would need to find the error covariance matrix $\mathbf{P}_{i}^{n}$ by evaluating $\mathrm{Cov}[\mathbf{X}_{i}^{n} - \mathbf{X}_{i}]$ using the estimator obtained here. This has not been done yet in this thesis so it is presented here.

We have the estimator $\mathbf{X}_{i}^{n}$ given as,

$$\mathbf{X}_{i}^{n} = \mathbf{X}_{i}^{i} + \mathbf{J}_{i}(\mathbf{X}_{i+1}^{n} - \mathbf{X}_{i+1}^{i}), \tag{4.215}$$

using the defined $\mathbf{J}_{i}$ from Section 4.4.3. Subtracting $\mathbf{X}_{i}$ from both sides and rearranging gives,

$$(\mathbf{X}_{i}^{n} - \mathbf{X}_{i}) - \mathbf{J}_{i}\mathbf{X}_{i+1}^{n} = (\mathbf{X}_{i}^{i} - \mathbf{X}_{i}) - \mathbf{J}_{i}\mathbf{X}_{i+1}^{i}. \tag{4.216}$$

Now, as stated in (Rauch et al., 1965), by taking the covariance of this and using the following facts:

$$\mathrm{E}[(\mathbf{X}_{i}^{n} - \mathbf{X}_{i})\mathbf{X}_{i+1}^{n}{}^{T}] = \mathrm{E}[(\mathbf{X}_{i}^{i} - \mathbf{X}_{i})\mathbf{X}_{i}^{i^{T}}] = \mathbf{0} \tag{4.217}$$

$$\mathrm{Cov}[\mathbf{X}_{i}^{i}] = \mathrm{Cov}[\mathbf{X}_{i}] - \mathbf{P}_{i}^{i} \tag{4.218}$$

$$\mathrm{Cov}[\mathbf{X}_{i+1}^{n}] = \mathrm{Cov}[\mathbf{X}_{i+1}] - \mathbf{P}_{i+1}^{n}, \tag{4.219}$$

we readily find the smoothing recurrence relation for $\mathbf{P}_{i}^{n}$. This will be shown now before Equations 4.217, 4.218, and 4.219 are proven. Equation 4.217 is used to infer that $(\mathbf{X}_{i}^{n} - \mathbf{X}_{i})$ is uncorrelated to $\mathbf{X}_{i+1}^{n}$ and that $(\mathbf{X}_{i}^{i} - \mathbf{X}_{i})$ is uncorrelated to $\mathbf{X}_{i}^{i}$. This is actually only true under the assumption that the prior exception $\mathrm{E}[\mathbf{X}_{0}]$ is chosen to be zero. Recall that if the prior expectation is chosen to be zero then the expectation of $\mathbf{X}_{i}$ and all estimators of $\mathbf{X}_{i}$ are zero (for any i), and therefore $\mathrm{E}[(\mathbf{X}_{i}^{n} - \mathbf{X}_{i})\mathbf{X}_{i+1}^{n}{}^{T}]$ and $\mathrm{E}[(\mathbf{X}_{i}^{n} - \mathbf{X}_{i})\mathbf{X}_{i+1}^{n}{}^{T}]$ are equal to the covariances $\mathrm{Cov}[\mathbf{X}_{i}^{n} - \mathbf{X}_{i}, \mathbf{X}_{i+1}^{n}{}^{T}]$ and $\mathrm{Cov}[\mathbf{X}_{i}^{n} - \mathbf{X}_{i}, \mathbf{X}_{i+1}^{n}{}^{T}]$ since $\mathrm{Cov}[\mathbf{A}, \mathbf{B}] = \mathrm{E}[\mathbf{AB}] - \mathrm{E}[\mathbf{A}]\mathrm{E}[\mathbf{B}]^{T}$ for any random vectors $\mathbf{A}$ and $\mathbf{B}$. So, with Equation 4.217, we see that we can write the following:

$$\text{Cov}[(\mathbf{X}_i^n - \mathbf{X}_i) - \mathbf{J}_i\mathbf{X}_{i+1}^n] = \text{Cov}[(\mathbf{X}_i^i - \mathbf{X}_i) - \mathbf{J}_i\mathbf{X}_{i+1}^i] \tag{4.220}$$

$$\text{Cov}[\mathbf{X}_i^n - \mathbf{X}_i] + \text{Cov}[\mathbf{J}_i\mathbf{X}_{i+1}^n] = \text{Cov}[\mathbf{X}_i^i - \mathbf{X}_i] + \text{Cov}[\mathbf{J}_i\mathbf{X}_{i+1}^i]. \tag{4.221}$$

Then, recalling that $\mathbf{X}_{i+1}^i = \mathbf{G}_{i+1}^i\mathbf{X}_i^i$, and using Equations 4.218, and 4.219, we have,

$$\mathbf{P}_i^n + \mathbf{J}_i\text{Cov}[\mathbf{X}_{i+1}^n]\mathbf{J}_i^T = \mathbf{P}_i^i + \mathbf{J}_i\mathbf{G}_{i+1}^i\text{Cov}[\mathbf{X}_i^i]{\mathbf{G}_{i+1}^i}^T\mathbf{J}_i^T \tag{4.222}$$

$$\mathbf{P}_i^n + \mathbf{J}_i(\text{Cov}[\mathbf{X}_{i+1}] - \mathbf{P}_{i+1}^n)\mathbf{J}_i^T = \mathbf{P}_i^i + \mathbf{J}_i\mathbf{G}_{i+1}^i(\text{Cov}[\mathbf{X}_i] - \mathbf{P}_i^i){\mathbf{G}_{i+1}^i}^T\mathbf{J}_i^T. \tag{4.223}$$

For $\text{Cov}[\mathbf{X}_{i+1}]$, recalling from the evolution equation that $\text{Cov}[\mathbf{X}_{i+1}] = \mathbf{G}_{i+1}\text{Cov}[\mathbf{X}_i]\mathbf{G}_{i+1}^T + \mathbf{W}_{i+1}$, and plugging this in gives,

$$\mathbf{P}_i^n + \mathbf{J}_i(\mathbf{G}_{i+1}\text{Cov}[\mathbf{X}_i]\mathbf{G}_{i+1}^T + \mathbf{W}_{i+1} - \mathbf{P}_{i+1}^n)\mathbf{J}_i^T = \mathbf{P}_i^i + \mathbf{J}_i\mathbf{G}_{i+1}^i(\text{Cov}[\mathbf{X}_i] - \mathbf{P}_i^i){\mathbf{G}_{i+1}^i}^T\mathbf{J}_i^T. \tag{4.224}$$

Simplifying this down gives,

$$\mathbf{P}_i^n = \mathbf{P}_i^i + \mathbf{J}_i\mathbf{P}_{i+1}^n\mathbf{J}_i^T - \mathbf{J}_i(\mathbf{G}_{i+1}^i\mathbf{P}_i^i{\mathbf{G}_{i+1}^i}^T + \mathbf{W}_{i+1})\mathbf{J}_i^T. \tag{4.225}$$

Then, recalling that $\mathbf{P}_{i+1}^i = \mathbf{G}_{i+1}^i\mathbf{P}_i^i{\mathbf{G}_{i+1}^i}^T + \mathbf{W}_{i+1}$ we finally have,

$$\mathbf{P}_i^n = \mathbf{P}_i^i + \mathbf{J}_i(\mathbf{P}_{i+1}^n - \mathbf{P}_{i+1}^i)\mathbf{J}_i^T, \tag{4.226}$$

which is indeed the correct smoothing recurrence relation for $\mathbf{P}_i^n$ we have given previously.

Now, we will prove Equations 4.217, 4.218, and 4.219 that were used to obtain this result. The work of (Rauch et al., 1965) states that the first equation can be verified by a lengthy manipulation of both the filtering and smoothing recurrence relations together. Then, we assume that they find the other two equations just as a consequence of the first. However, we can very easily verify these three equations instead with orthogonal projections of random variables, which was developed in Section 4.5.4 and Appendix Q of this thesis for what we called "Kalman's justification". It is stated in this Section 4.5.4 and proven in Appendix Q that the orthogonal projection of $\mathbf{X}_i$ onto the $\boldsymbol{\Upsilon}_n$ vector space (the vector space formed by the

DLM random vectors $\mathbf{Y}_1, ..., \mathbf{Y}_n$) is the conditional mean $\mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{Y}}_n]$, which is the estimator $\mathbf{X}_i^n$. Again, the only additional assumption needed here is that the prior expectation $\mathrm{E}[\mathbf{X}_0]$ is zero. Therefore, with the trick of Equation 4.138 we can write $\mathbf{X}_i$ as,

$$\mathbf{X}_i = \mathbf{X}_i^n + \tilde{\mathbf{X}}_i^n, \tag{4.227}$$

where $\tilde{\mathbf{X}}_i^n$ is defined as the random vector that contains the components of the random variables in $\mathbf{X}_i$ that are orthogonal to the $\boldsymbol{\Upsilon}_n$ vector space. Recall that this implies that $\mathrm{E}[\tilde{\mathbf{X}}_i^n \mathbf{X}_i^{nT}] = 0$, because all the elements of $\mathbf{X}_i^n$ and $\tilde{\mathbf{X}}_i^n$ are orthogonal to each other.

Now, using this we will show that $\mathrm{E}[(\mathbf{X}_i^i - \mathbf{X}_i)\mathbf{X}_i^{iT}] = \mathbf{0}$. We notice that we can write Equation 4.227 for the case $n = i$ as $\mathbf{X}_i = \mathbf{X}_i^i + \tilde{\mathbf{X}}_i^i$, so we have,

$$\begin{aligned}
\mathrm{E}[(\mathbf{X}_i^i - \mathbf{X}_i)\mathbf{X}_i^{iT}] &= \mathrm{E}[(\mathbf{X}_i^i - \mathbf{X}_i^i - \tilde{\mathbf{X}}_i^i)\mathbf{X}_i^{iT}] && (4.228)\\
&= \mathrm{E}[\tilde{\mathbf{X}}_i^i \mathbf{X}_i^{iT}] && (4.229)\\
&= \mathbf{0}, && (4.230)
\end{aligned}$$

which proves first part of Equation 4.217. Similarly for the second part of the equation, we find,

$$\begin{aligned}
\mathrm{E}[(\mathbf{X}_i^n - \mathbf{X}_i)\mathbf{X}_{i+1}^{n\,T}] &= \mathrm{E}[(\mathbf{X}_i^n - \mathbf{X}_i^n - \tilde{\mathbf{X}}_i^n)\mathbf{X}_{i+1}^{n\,T}] && (4.231)\\
&= \mathrm{E}[\tilde{\mathbf{X}}_i^n \mathbf{X}_{i+1}^{n\,T}] && (4.232)\\
&= \mathbf{0}. && (4.233)
\end{aligned}$$

For proving Equation 4.218 now, we start by writing,

$$\begin{aligned}
\mathbf{P}_i^i &= \mathrm{E}[(\mathbf{X}_i^i - \mathbf{X}_i)(\mathbf{X}_i^i - \mathbf{X}_i)^T] && (4.234)\\
&= \mathrm{E}[\mathbf{X}_i^i \mathbf{X}_i^{iT} + \mathbf{X}_i \mathbf{X}_i^T - \mathbf{X}_i \mathbf{X}_i^{iT} - \mathbf{X}_i^i \mathbf{X}_i^T] && (4.235)\\
&= \mathrm{Cov}[\mathbf{X}_i^i] + \mathrm{Cov}[\mathbf{X}_i] - \mathrm{E}[\mathbf{X}_i \mathbf{X}_i^{iT}] - \mathrm{E}[\mathbf{X}_i^i \mathbf{X}_i^T]. && (4.236)
\end{aligned}$$

And, with our orthogonality tricks, we find that,

$$\mathrm{E}[\mathbf{X}_i^i \mathbf{X}_i^T] = \mathrm{E}[\mathbf{X}_i^i (\mathbf{X}_i^i + \tilde{\mathbf{X}}_i^i)^T] \tag{4.237}$$

$$= \mathrm{E}[\mathbf{X}_i^i \mathbf{X}_i^i] \tag{4.238}$$

$$= \mathrm{Cov}[\mathbf{X}_i^i]. \tag{4.239}$$

This means also that we have $\mathrm{E}[\mathbf{X}_i \mathbf{X}_i^{iT}] = \mathrm{E}[\mathbf{X}_i^i \mathbf{X}_i^T]$ (since covariance matrices are symmetric). Plugging this into Equation 4.236, we see that we have,

$$\mathrm{Cov}[\mathbf{X}_i^i] = \mathrm{Cov}[\mathbf{X}_i] - \mathbf{P}_i^i, \tag{4.240}$$

which proves Equation 4.218. Equation 4.219 can be proved in fundamentally the same way as this one. To our knowledge, these three equations have never been proved in this way before.

4.6 Model Specification

So far we have seen how a DLM is specified by defining the model matrices $\mathbf{F}_i, \mathbf{G}_i, \mathbf{W}_i, \mathbf{V}_i$ at each index i. We have also seen two examples of DLMs (the multiple regression DLM and the local level DLM) and the recurrence relations used to estimate the states of the model given observed data. In this section, we will see how DLMs can be constructed by combining two or more DLMs together (West and Harrison, 1997). This is commonly done in practice for time series analysis since each individual DLM captures a different feature of the time series. For example, if a time series is known to depend upon some regressor variables and also has a smoothly varying background level, the modeler may wish to combine the multiple regression DLM with the local level DLM. For such a time series this should be better than just using one of the DLMs. Using the terminology of (West and Harrison, 1997), we refer to this as the superposition of DLMs.

The superposition of DLMs is done as follows. Let us denote m DLMs for data of the same dimension as DLM_j $\{\mathbf{F}_{ij}, \mathbf{G}_{ij}, V_{ij}, \mathbf{W}_{ij}\}$ for $j = 1, ..., m$ with state vector random vectors denoted as $\mathbf{X}_{ij}$ for $j = 1, ..., m$. The superposed DLM constructed from these DLMs is defined by the following block matrices:

$$\mathbf{F}_i = \begin{bmatrix} \mathbf{F}_{i1} & \mathbf{F}_{i2} & \dots & \mathbf{F}_{im} \end{bmatrix}, \tag{4.241}$$

$$\mathbf{G}_i = \begin{bmatrix} \mathbf{G}_{i1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{i2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{G}_{im} \end{bmatrix}, \tag{4.242}$$

and

$$\mathbf{W}_i = \begin{bmatrix} \mathbf{W}_{i1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{i2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{W}_{im} \end{bmatrix}, \tag{4.243}$$

and the new random vector representing the state vector at index i is

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{X}_{i1} \\ \mathbf{X}_{i2} \\ \vdots \\ \mathbf{X}_{im} \end{bmatrix}. \tag{4.244}$$

The form of $\mathbf{G}_i$ and $\mathbf{W}_i$ are what are called block diagonal matrices. We note that the model matrix $\mathbf{V}_i$ should be treated independently of this because it is defined based on the knowledge of the accuracy of the data alone.

As an example, we show the DLM model matrices for the superposition of the multiple regression DLM and the local level DLM. They are:

$$\mathbf{F}_i = \begin{bmatrix} 1 & \boldsymbol{\varphi}_i^T \end{bmatrix}, \tag{4.245}$$

$$\mathbf{G}_i = \mathbf{I}, \tag{4.246}$$

and

$$\mathbf{W}_i = \begin{bmatrix} W_{i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{i2} \end{bmatrix}, \tag{4.247}$$

where W_{i1} and $\mathbf{W}_{i2}$ are the $\mathbf{W}_i$ for the local level DLM and multiple regression DLM respectively, and the new random vector representing the state vector is

$$\mathbf{X}_i = \begin{bmatrix} M_i \\ \boldsymbol{\beta}_i \end{bmatrix}. \tag{4.248}$$

In the next sections, we specify four DLMs that can be superimposed to create a good model for a stratospheric ozone time series. These are also just four very common univariate DLMs that are fundamental to DLM time series analyses in general.

4.6.1 The Local Level Trend DLM

The second order polynomial DLM or "local level trend DLM" is specified by its model matrices as,

$$\mathbf{F}_i = \begin{bmatrix} 1 & 0 \end{bmatrix}, \tag{4.249}$$

$$\mathbf{G}_i = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \tag{4.250}$$

and

$$\mathbf{W}_i = \begin{bmatrix} W_{i1} & W_{i3} \\ W_{i3} & W_{i2} \end{bmatrix}. \tag{4.251}$$

Often this model is presented with its own unique observation and evolution equations rather than using the general DLM framework we have set up in this thesis. To show this, if we define the state vector as,

$$\mathbf{X}_i = \begin{bmatrix} M_i \\ A_i \end{bmatrix}, \tag{4.252}$$

the observation equation can be written as,

$$Y_i = M_i + \mathbf{v}_i, \qquad v \sim N[0, \mathbf{V}_i], \tag{4.253}$$

and the evolution equation can be written as two equations, rather than a single matrix equation, as,

$$M_i = M_{i-1} + A_{i-1} + w_{i1} \quad \text{and} \tag{4.254}$$

$$A_i = A_{i-1} + w_{i2}, \tag{4.255}$$

where,

$$\begin{bmatrix} w_{i1} \\ w_{i2} \end{bmatrix} \sim N[0, \mathbf{W}_i]. \tag{4.256}$$

The M_i component of the state vector is called the level component and the A_i component is called the trend component. Effectively, M_i represents the background level of the data, the same as we described for the local level DLM in Section 4.2.2. And, we can think of the second variable A_i as an addition to the basic local level DLM, and we see from the evolution equations that in a sense it is related to the change in the level. This is why it is called the trend component.

So, this DLM performs a similar function to the local level DLM in that it models a smoothly varying background level in the data. Lastly, we make the note that typically the modeller chooses W_{i3} to be 0. Then, the values of W_{i1} and W_{i2} alone control how quickly the background level fit can vary. With larger W_{i1} and W_{i2} allowing for more variation and smaller less variation. This is illustrated later in this thesis in Section 5.1.2.

4.6.2 The Multiple Regression DLM

The multiple regression DLM has already been introduced in a previous chapter as DLM $\{\boldsymbol{\varphi}_i^T, \mathbf{I}, \mathbf{V}_i, \mathbf{W}_i\}$, where $\boldsymbol{\varphi}_i$ is defined as,

$$\boldsymbol{\varphi}_i^T = \begin{bmatrix} x_{1i} & x_{2i} & \dots & x_{ki} \end{bmatrix} \qquad i = 1, 2, ..., n, \tag{4.257}$$

and the $x_{1i}, ..., x_{ki}$ are the values of each of the k regressors at index i.

This model allows for the capturing of the signals of each chosen regressor similarly to how this is done with an MLR model. The matrix $\mathbf{W}_i$ can be chosen to be anything, where the diagonal elements control the degree of variability in the estimation of the regression coefficients over the time series (independently of the other regressors), and the off-diagonal elements allow for the same control but allow for the variability to be coupled between regressors. Another possible choice is to specify $\mathbf{W}_i$ to be all zeros, allowing for no variability in the regression coefficients at all.

As stated earlier in this thesis, with this final choice of $\mathbf{W}_i$, the multiple regression DLM actually reduces to the equivalent of the MLR model. It is equivalent in form as well as leads to the same estimation of the regression coefficients. It can be seen from the evolution equation that $\boldsymbol{\beta}_i = \boldsymbol{\beta}_{i-1}$ in this case, and so if we define all $\boldsymbol{\beta}_i$ to then be $\boldsymbol{\beta}$, we may write the multiple regression DLM observation equation as $Y_i = \boldsymbol{\varphi}_i^T \boldsymbol{\beta} + e_i$, which is the MLR model equation. The next question becomes whether or not the DLM estimation procedure yields the same estimate of $\boldsymbol{\beta}$ as the MLR least squares estimate. Again, it turns out that it basically does. What happens is that the filtering recurrence relation becomes the same as a weighted RLS algorithm (see Appendix L for the weighted RLS algorithm). So, assuming the prior information used in the DLM is negligible, the last estimate ($\mathbf{x}_n^n$ if the time series has n data points) is identical to the GLS estimate with a diagonal covariance structure in the errors. This is intuitive because the multiple regression DLM requires specification of V_i, which is effectively the diagonal elements of the matrix we defined as $a^2\mathbf{V}$ for the GLS estimation theory. Further to this discussion about the filtering estimation, we also need to know what happens with the smoothing estimation. It is found that the smoothing recurrence relation reduces to $\mathbf{x}_i^n = \mathbf{x}_{i+1}^n$ in this case. So, this means that the sequence of smoothed estimates will have the same objects and they are equal to $\mathbf{x}_n^n$, the equivalent of the GLS estimate. Furthermore, we get essentially equivalent error estimation too because $\mathbf{P}_n^n$ is identical to the covariance matrix of the GLS estimator, and then for the smoothing recurrence relation we also get $\mathbf{P}_i^n = \mathbf{P}_{i+1}^n$. See Appendix T for all these details.

4.6.3 Autocorrelation DLMs

4.6.3.1 First Order

A first order autoregressive model, AR(1), (which was introduced in Section 2.4.3) can be formulated as a DLM with the following model matrices:

$$\mathbf{F}_i = \begin{bmatrix} 1 \end{bmatrix}, \qquad \mathbf{G}_i = \begin{bmatrix} \rho \end{bmatrix}, \qquad \text{and} \quad \mathbf{W}_i = \begin{bmatrix} \sigma_{AR}^2 \end{bmatrix}, \tag{4.258}$$

where we define σ_{AR}^2 as the variance of $\mathbf{w}_i$ in the observation equation. So, the observation and evolution equations take the following form:

$$Y_i = X_i + v_i, \qquad v_i \sim N[0, V_i] \tag{4.259}$$

$$X_i = \rho X_{i-1} + w_i, \qquad w_i \sim N[0, \sigma_{AR}^2]. \tag{4.260}$$

The evolution equation takes the same form of an AR(1) model equation (see Equation 2.31), and the observation equation tells us that the data is only the AR signal plus some noise. So, this is effectively an AR(1) model where the error in the model and the data have been separated into two equations, and nicely into the standard DLM form so that we can use it as a DLM.

4.6.3.2 Order p

Auto regression models of any order p, AR(p), (which has not been given in this thesis yet) can be expressed in the DLM form with the following model matrices:

$$\mathbf{F}_i = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}, \tag{4.261}$$

$$\mathbf{G}_i = \begin{bmatrix} \rho_1 & \rho_2 & \rho_3 & \dots & \rho_p \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}, \tag{4.262}$$

and

$$\mathbf{W}_i = \begin{bmatrix} \sigma_{AR}^2 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}, \tag{4.263}$$

where the random variable that represents the state vector is

$$\mathbf{X}_i = \begin{bmatrix} Z_i \\ Z_{i-1} \\ \vdots \\ Z_{i-(p-1)} \end{bmatrix}. \tag{4.264}$$

For a brief explanation, we see that by putting these matrices into the general evolution equation form of

$$\mathbf{X}_i = \mathbf{G}_i \mathbf{X}_{i-1} + \mathbf{w}_i, \qquad \mathbf{w}_i \sim N[0, \mathbf{W}_i], \tag{4.265}$$

that the first element of the evolution equation is the same as the AR(p) model equation and the other elements are congruent but non-informative. The AR(p) model equation has not been shown thus far in this thesis, but it is indeed given by what we would find from this first element of the evolution equation. The AR(p) model equation is given as,

$$Z_i = \sum_{j=1}^{p} \rho_i Z_{i-j} + \epsilon_i, \qquad \epsilon_i \sim N[0, \sigma_{AR}^2]. \tag{4.266}$$

The other elements 2 through p of the evolution equation simply state the following:

$$Z_{i-1} = Z_{i-1}, \quad Z_{i-2} = Z_{i-2}, \quad \dots, \quad \text{and} \quad Z_{i-(p-1)} = Z_{i-(p-1)}. \tag{4.267}$$

So, the evolution equation is a valid matrix equation for an AR(p) process. Then, the observation equation is the same as it was for the AR(1) case, simply the AR signal plus some noise. Lastly, we should also note that the AR(p) DLM can be formulated in other ways. One such way is actually the same as what is presented here but with $\mathbf{G}_i$ transposed.

4.6.4 The Fourier Form Seasonal DLM

The Fourier form seasonal DLM is defined by the following model matrices:

$$\mathbf{F}_i = \begin{bmatrix} 1 & 0 \end{bmatrix}, \tag{4.268}$$

$$\mathbf{G}_i = \begin{bmatrix} cos(\omega) & sin(\omega) \\ -sin(\omega) & cos(\omega) \end{bmatrix}, \tag{4.269}$$

and

$$\mathbf{W}_i = \begin{bmatrix} \sigma^2_{seas} & 0 \\ 0 & \sigma^2_{seas} \end{bmatrix}. \tag{4.270}$$

The rest of this section will be an explanation of these model matrices. A discrete sinusoid with a frequency ω can be written as,

$$s[i] = s_i = acos(\omega i) + bsin(\omega i), \tag{4.271}$$

and a similar quantity can be defined as,

$$s[i]^* = s^*_i = -asin(\omega i) + bcos(\omega i). \tag{4.272}$$

It can easily be verified using trigonometry identities that

$$s_i = s_{i-1}cos(\omega) + s^*_{i-1}sin(\omega), \tag{4.273}$$

and similarly that

$$s^*_i = -s_{i-1}sin(\omega) + s^*_{i-1}cos(\omega). \tag{4.274}$$

These equations give a way of calculating the values of s_i and s^*_i given the values at the previous index $i-1$. The point of this is that this is the way the DLM evolution equation works, so this can be used to construct a DLM evolution equation. More specifically, the evolution equation is in the following form:

$$\begin{bmatrix} S_i \\ S_i^* \end{bmatrix} = \begin{bmatrix} cos(\omega) & sin(\omega) \\ -sin(\omega) & cos(\omega) \end{bmatrix} \begin{bmatrix} S_{i-1} \\ S_{i-1}^* \end{bmatrix} + \mathbf{w}_i, \qquad \mathbf{w}_i \sim N[0, \mathbf{W}_i], \tag{4.275}$$

where we use capital S's now to denote that these are random variables. We see that the random vector that represents the state vector in this evolution equation is

$$\mathbf{X}_i = \begin{bmatrix} S_i \\ S_i^* \end{bmatrix}. \tag{4.276}$$

For the $\mathbf{W}_i$ model matrix, it makes sense considering the symmetry of these equations to have it be a diagonal matrix with equal diagonal elements. The defined model matrices at the beginning of this section reflect this evolution equation.

Now, we choose $\mathbf{F} = \begin{bmatrix} 1 & 0 \end{bmatrix}$ so that the observation equation is described by the S_i signal as,

$$Y_i = S_i + v_i, \qquad v_i \sim N[0, V_i]. \tag{4.277}$$

This covers all the model matrices discussed. Again, $\mathbf{V_i}$ is defined independently based on the error in the data. Lastly, it is important to note that if σ^2_{seas} is chosen to be zero, then the smoothed estimate of S_i (and S_i^* technically) for this DLM would be sinusoidal with constant amplitude. If σ^2_{seas} is non-zero then the smoothed estimate is sinusoidal with a variable amplitude, with the degree of variability depending on σ^2_{seas}.

4.6.5 Superposition Example for a Stratospheric Ozone Time Series

In this section, we use the DLMs presented in the last four sections and the DLM superposition principle to create a DLM that is suitable for modelling a stratospheric ozone time series. But first, we go through the details of the application of each of the individual DLMs that we use.

For the local level trend DLM, since the values of W_{i1} and W_{i2} basically control the same thing, the smoothness of the background level fit, it is acceptable to use $W_{i1} = 0$ and have W_{i2} as the only non zero element of the $\mathbf{W}_i$ matrix. With this, something that is fairly useful

for trend analysis happens. Consider the sequence of smoothed estimates for a local level trend DLM. It turns out that for this choice of $\mathbf{W}_i$ the trend component of this sequence is equal to the forward difference of the level component of this sequence (see Appendix U for the proof). So, the trend component is the numerical derivative of the level component. This is important because we use this fact later in this thesis when we show a full picture of the state of ozone trends in the stratospheric with the SOO data record using the DLM procedure. We will define W_{i2} in the local level trend DLM as σ^2_{trend} where we loosely call this variable the "trend variance".

For the regressors in the multiple regression DLM, we use the ENSO, SOLAR, QBOA, QBOB, and AOD indexes described in Section 2.2. We also choose the $\mathbf{W}_i$ matrix for this DLM to be all zeros so that the regression coefficient estimates are constant over the time series like they are in the MLR model. We also use the AR(1) DLM as presented in Section 4.6.3.1. In the stratospheric ozone trend literature, both the AR(1) and AR(2) DLMs have been used before.

Lastly, we use two Fourier form seasonal DLMs. One with a full-year period ($\omega = \frac{2\pi}{12}$) and one with a half-year period ($\omega = \frac{2\pi}{6}$). With monthly data these ω's have the desired periods of 12 and 6. For a note about using this DLM in this context, recall the discussion of choosing σ^2_{seas} at the end of Section 4.6.4. By choosing σ^2_{seas} to be zero for the full-year period DLM, the seasonal cycle of ozone data is modelled as a sinusoid, but, by choosing it to be non-zero we allow for the model to pick up variable amplitude from year to year in the seasonal cycle.

Now, combining all these models with the DLM superposition principal, we have an appropriate DLM for a monthly mean stratospheric ozone time series defined by the following

model matrices:

$$\mathbf{F}_i^T = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ ENSO_i \\ SOLAR_i \\ QBOA_i \\ QBOB_i \\ AOD_i \\ 1 \end{bmatrix} \tag{4.278}$$

where we define $REGRESSOR_i$ as the value of the regressors time series index at index i,

$$\mathbf{G}_i = \begin{bmatrix} \mathbf{G}_{1i} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{2i} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_{3i} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{G}_{4i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{G}_{5i} \end{bmatrix}, \tag{4.279}$$

where

$$\mathbf{G}_{1i} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \tag{4.280}$$

$$\mathbf{G}_{2i} = \begin{bmatrix} cos\left(\frac{2\pi}{12}\right) & sin\left(\frac{2\pi}{12}\right) \\ -sin\left(\frac{2\pi}{12}\right) & cos\left(\frac{2\pi}{12}\right) \end{bmatrix}, \tag{4.281}$$

$$\mathbf{G}_{3i} = \begin{bmatrix} cos\left(\frac{2\pi}{6}\right) & sin\left(\frac{2\pi}{6}\right) \\ -sin\left(\frac{2\pi}{6}\right) & cos\left(\frac{2\pi}{6}\right) \end{bmatrix}, \tag{4.282}$$

$$\mathbf{G}_{4i} = \mathbf{I}_{5x5}, \tag{4.283}$$

and

$$\mathbf{G}_{5i} = \left[\rho\right], \tag{4.284}$$

and

$$diag(\mathbf{W}_i) = \begin{bmatrix} 0 \\ \sigma^2_{trend} \\ \sigma^2_{seas,1} \\ \sigma^2_{seas,1} \\ \sigma^2_{seas,2} \\ \sigma^2_{seas,2} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \sigma^2_{AR} \end{bmatrix}, \tag{4.285}$$

where the off-diagonals of $\mathbf{W}_i$ are all zeros and we differentiate the two seasonal components variance with subscripts 1 and 2. With this formulation, the random vector that represents the state vector is

$$\mathbf{X}_i = \begin{bmatrix} M_i \\ A_i \\ S_{i,1} \\ S^*_{i,1} \\ S_{i,2} \\ S^*_{i,2} \\ \boldsymbol{\beta}_i \\ X_{i,AR} \end{bmatrix}, \tag{4.286}$$

where we define the random variable representing the state vector of the AR component as $X_{i,AR}$. Lastly, we note that if the monthly mean ozone data we are modelling is of relative

anomalies or anything deseasonalized then this DLM without the Fourier form DLMs can be sufficient since the seasonal cycle is already considered to be removed from the data.

Now, something that should be alarming is noticing that the model presented requires specification of the variables $\rho, \sigma_{trend}, \sigma_{seas,1}, \sigma_{seas,2}$, and σ_{AR}. Luckily, there exist procedures for estimating these parameters from the data, which we will get into shortly. In fact, rather than just estimating these parameters, we use MCMC algorithms to draw a sample from their probability distribution given the observed data. We can then run the DLM estimate for each sample point, allowing for the capability to obtain statistical inferences about stratospheric ozone taking into account the uncertainty of these parameters in a Monte Carlo fashion. This is the subject of Section 5.1 in the next chapter and is essentially the defining feature of the developed DLM procedure of this thesis work.

4.6.6 Selection of the DLM Prior

As stated previously, typically the DLM prior values of $\mathbf{x}_0^0$ and $\mathbf{P}_0^0$ are chosen to be $\mathbf{x}_0^0 = \mathbf{0}$ and $\mathbf{P}_0^0 = \kappa\mathbf{I}$ where κ is a large number. This selection of large diagonals on $\mathbf{P}_0^0$ leads to the prior estimate $\mathbf{x}_0^0$ having little effect on the estimation of the states after the first few indices. But, as stated when we specified the Kalman filter recurrence relations in Section 4.4.2, these prior values could be almost anything. If something is known about the state vector beforehand then that information can be incorporated into $\mathbf{x}_0^0$ and $\mathbf{P}_0^0$, where $\mathbf{x}_0^0$ is the best guess of the state vector and $\mathbf{P}_0^0$ expresses the degree of uncertainty in this guess.

For the ozone model in Section 4.6.5 we will take the first approach and set $\mathbf{x}_0^0 = \mathbf{0}$ and have $\mathbf{P}_0^0$ large (except for the caveat which is discussed in the paragraph below). The guess of $\mathbf{0}$ is actually a reasonable guess for all the SOO altitude-latitude regions since the regression coefficients, the initial background level, trend, and the AR(1) terms could all be on either side of zero. For $\mathbf{P}_0^0$, there is a limit to how large the diagonals can be until it starts causing problems. The problem it causes is the variance of the first few estimates can become abnormally large. But, there is often a wide region of values where the results are acceptable and all identical to each other. It is important to find this region and use a reasonable $\mathbf{P}_0^0$ because this can also affect the probability distributions of the DLM's unknown parameters (i.e. $\rho, \sigma_{trend}, \sigma_{seas,1}, \sigma_{seas,2}$, and σ_{AR} in the DLM presented in the previous section).

For the last diagonal of $\mathbf{P}_0^0$ in the ozone DLM, the one corresponding to the AR(1) part, we find that selection of it is very sensitive to making the first few variance estimates abnormally large and causing the problem discussed. Now, recall that for the AR(1) model, we derived the variance to be $\frac{\sigma_{AR}^2}{1-\rho^2}$ in Equation 2.35. The variance estimates $\mathbf{P}_i^n$, whether they are filtering, smoothing, or prediction, for the AR(1) DLM described in Section 4.6.3 are not the same as this, as we might be inclined to believe. This is because these variance estimates are conditioned on the observed data, whereas this was not part of the setup when we derived Equation 2.35. However, if we have $V_i >> \sigma_{AR}^2$, then it is the case that the variance estimates $\mathbf{P}_i^n$ are equal to $\frac{\sigma_{AR}^2}{1-\rho^2}$. For stratospheric ozone data, V_i is typically a significant amount greater than σ_{AR}^2, so this turns out to be a good estimate for all the $\mathbf{P}_i^n$ in this application. So, for the last diagonal element of $\mathbf{P}_0^0$ for the ozone model we will actually use this $\frac{\sigma_{AR}^2}{1-\rho^2}$ quantity. In practice, it is found that this works well for all the altitude-latitude regions of the SOO data record, not causing any problems with the first few estimates. Furthermore, as expected, the filtered estimates and smoothed estimates for the variance of the AR(1) term of the ozone DLM remain close to this value as they are sequentially calculated with the recurrence relations. Again, this is expected because V_i is a fair amount greater than σ_{AR}^2.

4.7 The Model Fit

In this section, we give the model fit for the DLM, like how we gave the model fit for the MLR model in Section 2.4.4. The contents of this section, of what we choose the model fit to be, are hopefully intuitive by now.

Consider observing data $\mathbf{y}_1, ..., \mathbf{y}_n$, or just as well consider observing one-dimensional data $y_1, .., y_n$ like the case of just a single time series. The model fit that we construct from our DLM state vector estimates is simply $\mathbf{F}_i\mathbf{x}_i^n$ at each index i. Of course, it makes sense that we use the smoothed estimate $\mathbf{x}_i^n$ here rather than say the filtered estimate $\mathbf{x}_i^i$, the one-step-ahead prediction estimate $\mathbf{x}_i^{i-1}$, or any other estimate, since the smoothed estimate takes the most information into account. If we define the model fit estimator at index i given observed data until index n as $\hat{\mathbf{Y}}_i^n$, we can write it as,

$$\hat{\mathbf{Y}}_i^n = \mathbf{F}_i \hat{\mathbf{X}}_i^n, \tag{4.287}$$

where we define $\hat{\mathbf{X}}_i^n$ as the Gaussian distributed random vector with mean $\mathbf{x}_i^n$ and covariance $\mathbf{P}_i^n$. Taking the covariance, we have the covariance of the model fit given as,

$$\text{Cov}[\hat{\mathbf{Y}}_i^n] = \mathbf{F}_i \mathbf{P}_i^n \mathbf{F}_i^T. \tag{4.288}$$

Notice that $\hat{\mathbf{Y}}_i^n$ is Gaussian since it is a linear function of a Gaussian distributed random vector. So, the probability distribution of the model fit estimator is Gaussian with a mean of $\mathbf{F}_i \mathbf{x}_i^n$ and covariance of $\mathbf{F}_i \mathbf{P}_i^n \mathbf{F}_i^T$.

4.7.1 Modelling Ozone

Consider the DLM for stratospheric ozone presented in Section 4.6.5 without the seasonal DLM components. We will write the approximation (or model fit) of this DLM in a way that is analogous to what was written in Equation 2.2 for the MLR model. We have,

$$\begin{aligned} OZONE(t) = a_1 QBOA(t) + a_2 QBOB(t) + a_3 SOLAR(t) + a_4 ENSO(t) + a_5 AOD(t) \\ + LEVEL(t) + AR1(t), \end{aligned} \tag{4.289}$$

where $LEVEL(t)$ and $AR1(t)$ are the background level and AR(1) components of the sequence of smoothed estimates, and the coefficients $a_1, ..., a_5$ are the regression coefficient smoothed estimates, which are the same for all indices (as discussed in Section 4.6.5) so we state them here as a single number rather than a function of time. This is what the DLM model fit looks like for the ozone model application we are developing.

In this chapter, we have provided introductory buildup for the DLM, shown the recurrence relations used to estimate a DLM's states (along with the theory behind them), and explained how to construct DLMs suitable for time series analysis. In the next chapter, the developed DLM procedure of this thesis work will finally be described.

5 The Dynamic Linear Model Procedure

In this chapter, we describe the DLM procedure we have developed for quantifying trends in time series. Likewise to the Prais-Winsten estimation, what we have is a procedure that optimizes a model, in this case, the DLM, and also does a better job in the error analysis. We also note here that this is a procedure whose application with a computer becomes fairly computationally expensive, unlike the Prais-Winsten estimation.

In Section 5.1 we discuss the most defining topic for the DLM procedure of how to estimate required input parameters to the DLM. This requires MCMC experiments. For this, some additional DLM theory about the DLM likelihood function is needed. In Section 5.2 some final details of the DLM procedure are described and we give a high-level overview of how the procedure is executed. Lastly, in Section 5.3 we summarize all of the parameters for the DLM procedure that the modeller must specify.

5.1 Estimating Unknown DLM Parameters

In this section, we show how we can estimate the unknown parameters of a DLM. We will denote the unknown DLM parameters as $\boldsymbol{\theta}$. For example, for the DLM in Section 4.6.5 we have $\boldsymbol{\theta} = \begin{bmatrix} \rho & \sigma_{trend} & \sigma_{seas,1} & \sigma_{seas,2} & \sigma_{AR} \end{bmatrix}$. What we are interested in is the probability distribution of $\boldsymbol{\theta}$ given the observed data. We write this as $p(\boldsymbol{\theta}|\tilde{\mathbf{y}}_n)$. This distribution can not be calculated directly, but the probability distribution of the observed data for a given DLM, written as $p(\tilde{\mathbf{y}}_n|\boldsymbol{\theta})$, can be. So, consider Bayes theorem with these distributions (see Section 4.5.2.1 for an introduction to Bayes theorem) as,

$$p(\boldsymbol{\theta}|\tilde{\mathbf{y}}_n) \propto p(\tilde{\mathbf{y}}_n|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{5.1}$$

Before we show how we calculate $p(\tilde{\mathbf{y}}_n|\boldsymbol{\theta})$ for a DLM, we will discuss what can be done with

the above Bayes formula to obtain information about $p(\boldsymbol{\theta}|\tilde{\mathbf{y}}_n)$. First, there exists numerical optimization algorithms to find maximums of functions when an analytical solution is not obtainable. These algorithms can be used to find the $\boldsymbol{\theta}$ that maximizes the right-hand side the equation, and hence simultaneously the left-hand side $p(\boldsymbol{\theta}|\tilde{\mathbf{y}}_n)$. The result is that the most probable $\boldsymbol{\theta}$ for the DLM given the observed data becomes known. Second, we can use an MCMC algorithm to obtain a sample from $p(\boldsymbol{\theta}|\tilde{\mathbf{y}}_n)$. The topic of MCMC was introduced in Chapter 3 of this thesis. Recall that for the Metropolis-Hastings MCMC algorithm, the only required knowledge about the probability distribution we wish to draw a sample from is the ability to evaluate a function that it is proportional too (i.e. $g(\mathbf{x})$ in the Chapter 3). This is exactly what we have in the right-hand side of Equation 5.1. So, to be clear, using the notation of Chapter 3 on MCMC we consider $p(\boldsymbol{\theta}|\tilde{\mathbf{y}}_n)$ to be the target distribution (or $p(\mathbf{x})$) and $p(\tilde{\mathbf{y}}_n|\boldsymbol{\theta})p(\boldsymbol{\theta})$ to be the function $g(\mathbf{x})$ we use in the Metropolis-Hastings algorithm, if we consider that $\mathbf{x} = \boldsymbol{\theta}$. Again, with this, we can obtain an approximate sample from $p(\boldsymbol{\theta}|\tilde{\mathbf{y}}_n)$, and as was hinted at previously in this thesis, this is useful for stratospheric ozone modelling, or any application for that matter, because it allows us to take into account the uncertainty of $\boldsymbol{\theta}$ in a Monte Carlo fashion by running the DLM for each of the $\boldsymbol{\theta}$ sample points. As for the prior distribution $p(\boldsymbol{\theta})$ on the right-hand side of the equation, it is typically specified in a subjective manner by the modeller. This is discussed more in Section 5.1.2.

5.1.1 The DLM Likelihood Function

In this section, the likelihood function $p(\tilde{\mathbf{y}}_n|\boldsymbol{\theta})$ for a DLM will be derived. The probability distribution of a sequence can be written as,

$$p(\tilde{\mathbf{y}}_n) = p(\mathbf{y}_1, ..., \mathbf{y}_n) = p(\mathbf{y}_1)p(\mathbf{y}_2|\mathbf{y}_1)p(\mathbf{y}_3|\mathbf{y}_2, \mathbf{y}_1)...p(\mathbf{y}_n|\mathbf{y}_{n-1}, ..., \mathbf{y}_1) \tag{5.2}$$

$$= p(\mathbf{y}_1)\prod_{j=2}^{n} p(\mathbf{y}_j|\mathbf{y}_1, ..., \mathbf{y}_{j-1}). \tag{5.3}$$

In the context of a sequence being modelled by a DLM, we calculate the probability of a sequence given the DLM model we have defined. So, if we wish to indicate this we may explicitly write this distribution conditioned on $\boldsymbol{\theta}$. Since $\boldsymbol{\theta}$ is the only thing that we anticipate

may change about the DLM model we are working with, writing it this way is intuitive. So, we have,

$$p(\tilde{\mathbf{y}}_n|\boldsymbol{\theta}) = p(\mathbf{y}_1|\boldsymbol{\theta}) \prod_{j=2}^{n} p(\mathbf{y}_j|\mathbf{y}_1, ..., \mathbf{y}_{j-1}, \boldsymbol{\theta}). \tag{5.4}$$

We will now find the distribution $p(\mathbf{y}_j|\mathbf{y}_1, ..., \mathbf{y}_{j-1}, \boldsymbol{\theta})$ (or $p(\mathbf{y}_j|\tilde{\mathbf{y}}_{j-1}, \boldsymbol{\theta})$ using our defined notation) for a DLM model. Given the DLM observation equation,

$$\mathbf{Y}_j = \mathbf{F}_j\mathbf{X}_j + \mathbf{v}_j, \qquad \mathbf{v}_j \sim N[0, \mathbf{V}_j], \tag{5.5}$$

the expectation and covariance of $\mathbf{Y}_j$ conditioned on $\tilde{\mathbf{y}}_{j-1}$ and $\boldsymbol{\theta}$ can be calculated as follows:

$$\begin{aligned} \mathrm{E}[\mathbf{Y}_j|\tilde{\mathbf{y}}_{j-1}, \boldsymbol{\theta}] &= \mathbf{F}_j\mathrm{E}[\mathbf{X}_j|\tilde{\mathbf{y}}_{j-1}, \boldsymbol{\theta}] + \mathrm{E}[\mathbf{v}_j|\tilde{\mathbf{y}}_{j-1}, \boldsymbol{\theta}] && (5.6)\\ &= \mathbf{F}_j\mathbf{x}_j^{j-1} + \mathrm{E}[\mathbf{v}_j] && (5.7)\\ &= \mathbf{F}_j\mathbf{x}_j^{j-1}, && (5.8) \end{aligned}$$

and

$$\begin{aligned} \mathrm{Cov}[\mathbf{Y}_j|\tilde{\mathbf{y}}_{j-1}, \boldsymbol{\theta}] &= \mathbf{F}_j\mathrm{Cov}[\mathbf{X}_j|\tilde{\mathbf{y}}_{j-1}, \boldsymbol{\theta}]\mathbf{F}_j^T + \mathrm{Cov}[\mathbf{v}_j|\tilde{\mathbf{y}}_{j-1}, \boldsymbol{\theta}] && (5.9)\\ &= \mathbf{F}_j\mathbf{P}_j^{j-1}\mathbf{F}_j^T + \mathrm{Cov}[\mathbf{v}_j] && (5.10)\\ &= \mathbf{F}_j\mathbf{P}_j^{j-1}\mathbf{F}_j^T + \mathbf{V}_j, && (5.11) \end{aligned}$$

where $\mathbf{x}_j^{j-1}$ and $\mathbf{P}_j^{j-1}$ are defined in Section 4.4.1. The conditioning on $\boldsymbol{\theta}$ which was not seen in Section 4.4 is nothing to be alarmed about since it is just stated to remind us that we are working with a given DLM. In prior sections of this thesis, we could say that conditioning on $\boldsymbol{\theta}$ or just on the given DLM was implied in all the work.

Also, $p(\mathbf{y}_j|\tilde{\mathbf{y}}_{j-1}, \boldsymbol{\theta})$ must be a Gaussian distribution. If we define the random variable that represents this distribution as $\mathbf{Y}_j^{j-1}$, then by definition we have,

$$\mathbf{Y}_j^{j-1} = \mathbf{F}_j\hat{\mathbf{X}}_j^{j-1} + v_j, \tag{5.12}$$

where $\hat{\mathbf{X}}_j^{j-1}$ was defined in Section 4.7. Now, it can be seen that since $\mathbf{Y}_j^{j-1}$ is a linear combination of Gaussian distributed random vectors it is also Gaussian. So, $p(\mathbf{y}_j|\tilde{\mathbf{y}}_{j-1}, \boldsymbol{\theta})$ is given as,

$$p(\mathbf{y}_j|\tilde{\mathbf{y}}_{j-1}, \boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^k|\mathbf{C}_j|}} e^{-\frac{1}{2}(\mathbf{y}_j - \mathbf{F}_j\mathbf{x}_j^{j-1})^T\mathbf{C}_j^{-1}(\mathbf{y}_j - \mathbf{F}_j\mathbf{x}_j^{j-1})}, \tag{5.13}$$

where we define $\mathbf{C}_j$ as $\text{Cov}[\mathbf{Y}_j|\tilde{\mathbf{y}}_{j-1}, \boldsymbol{\theta}]$ or $\mathbf{F}_j\mathbf{P}_j^{j-1}\mathbf{F}_j^T + \mathbf{V}_j$ and where k is the number of elements of $\mathbf{y}_j$. This equation actually also encompasses the $p(\mathbf{y}_1|\boldsymbol{\theta})$ case where there is no previous data. So, using this in Equation 5.4 we finally can write the likelihood function as,

$$p(\tilde{\mathbf{y}}_n|\boldsymbol{\theta}) = \prod_{j=1}^{n} \frac{1}{\sqrt{(2\pi)^k|\mathbf{C}_j|}} e^{-\frac{1}{2}(\mathbf{y}_j - \mathbf{F}_j\mathbf{x}_j^{j-1})^T\mathbf{C}_j^{-1}(\mathbf{y}_j - \mathbf{F}_j\mathbf{x}_j^{j-1})}. \tag{5.14}$$

An important thing to notice here is that this can be calculated along with the Kalman filter recurrence relation algorithm. We see this because the quantities $\mathbf{C}_j$ and $\mathbf{x}_j^{j-1}$ are inherently calculated by the Kalman filter algorithm. In this section, we have shown a calculable likelihood function for the DLM, which we can use in the Metropolis-Hastings MCMC algorithm to draw a sample from $p(\boldsymbol{\theta}|\tilde{\mathbf{y}}_n)$ for our DLM procedure.

5.1.1.1 A Note on this in Practice

Since we do not care about any multiplicative constant in $g(\mathbf{x})$ in the Metropolis-Hastings algorithm, let us write Equation 5.14 without the constant multiplier that does not depend on $\boldsymbol{\theta}$ as follows:

$$p(\tilde{\mathbf{y}}_n|\boldsymbol{\theta}) \propto \prod_{j=1}^{n} \frac{1}{\sqrt{|\mathbf{C}_j|}} e^{-\frac{1}{2}(\mathbf{y}_j - \mathbf{F}_j\mathbf{x}_j^{j-1})^T\mathbf{C}_j^{-1}(\mathbf{y}_j - \mathbf{F}_j\mathbf{x}_j^{j-1})}, \tag{5.15}$$

and define this right-hand side as $a_{n\boldsymbol{\theta}}$. Now, because this is typically an exponential with a large negative exponent it is not practical for a computer to directly calculate $a_{n\boldsymbol{\theta}}$, or more specifically, $a_{n\boldsymbol{\theta}'}/a_{n\boldsymbol{\theta}_{i-1}}$ in the Metropolis-Hastings algorithm. But, if we take the logarithm we have,

$$\ln(a_{n\boldsymbol{\theta}}) = -\sum_{j=1}^{n}[\ln(\sqrt{|\mathbf{C}_j|}) + \frac{1}{2}(\mathbf{y}_j - \mathbf{F}_j\mathbf{x}_j^{j-1})^T\mathbf{C}_j^{-1}(\mathbf{y}_j - \mathbf{F}_j\mathbf{x}_j^{j-1})], \tag{5.16}$$

which is practical to calculate. So, in the Metropolis-Hastings algorithm we calculate the ratio $a_{n\boldsymbol{\theta}'}/a_{n\boldsymbol{\theta}_{i-1}}$ with the formula $exp(\ln(a_{n\boldsymbol{\theta}'}) - \ln(a_{n\boldsymbol{\theta}_{i-1}}))$ instead, making it doable for a computer.

5.1.2 Prior Distribution Selection

In this section, we discuss the choice of the prior distribution $p(\boldsymbol{\theta})$. The most simple choice is a uniform distribution across the whole probability space so that essentially no prior information is incorporated. With this, the shape of $p(\boldsymbol{\theta}|\tilde{\mathbf{y}}_n)$ becomes the shape of $p(\tilde{\mathbf{y}}_n|\boldsymbol{\theta})$ since we would have $p(\boldsymbol{\theta}|\tilde{\mathbf{y}}_n) \propto p(\tilde{\mathbf{y}}_n|\boldsymbol{\theta})$.

If we instead choose to incorporate prior information, then it is often the case that the modeller would do so independently for each variable in the $\boldsymbol{\theta}$ vector. If this is done, the ratio $\frac{p(\boldsymbol{\theta}')}{p(\boldsymbol{\theta}_{i-1})}$ that shows up in the Metropolis-Hastings algorithm becomes the product of the individual prior distributions as,

$$\frac{p(\boldsymbol{\theta}')}{p(\boldsymbol{\theta}_{i-1})} = \prod_{j=1}^{n} \frac{p((\boldsymbol{\theta}')_j)}{p((\boldsymbol{\theta}_{i-1})_j)}, \tag{5.17}$$

where $(\boldsymbol{\theta})_j$ is the jth of n elements of $\boldsymbol{\theta}$. This is similar to the discussion that was had in Section 5.1.2 on the Metropolis-Hastings proposal distribution.

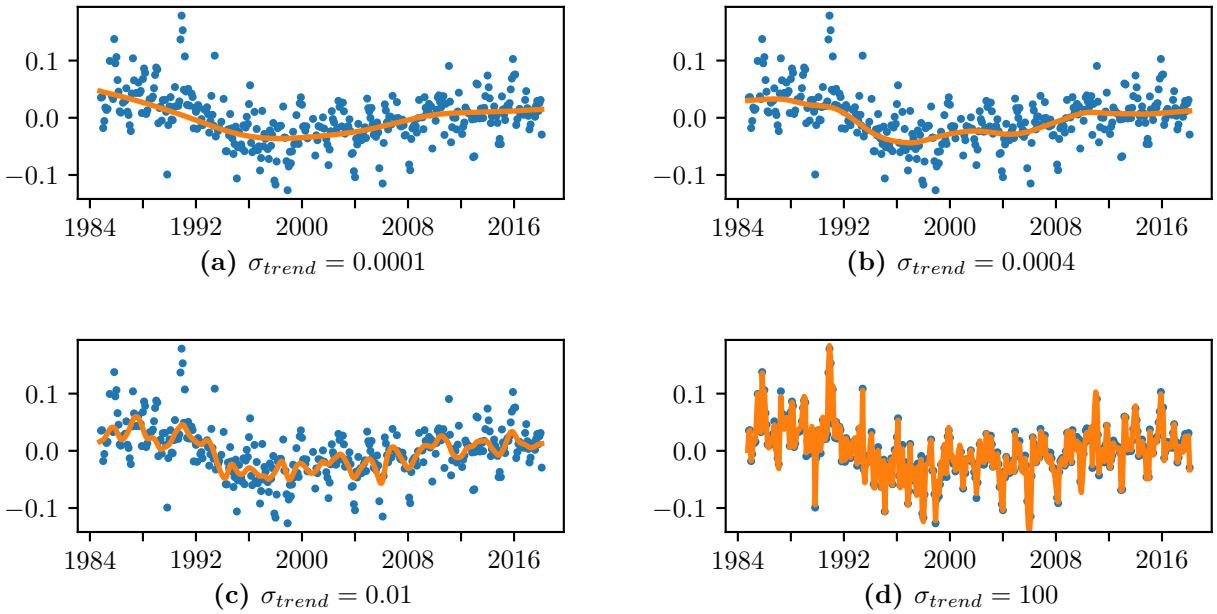
With regards to the ozone model of Section 4.6.5, notice that we have mostly standard deviation parameters in $\boldsymbol{\theta}$. For standard deviation parameters, we can at the very least select a uniform prior with bounds at 0 and positive infinity because standard deviations cannot be less than 0.

Now, we recommend that if there is reliable prior information available the modeller should not be scared to include it. Also, we note that using a prior distribution to constrain a parameter can sometimes event be necessary.

To illustrate this further, consider the σ_{trend} element of $\boldsymbol{\theta}$ for the ozone model in Section 4.6.5. Recall that we said earlier that this parameter essentially controls how quickly the background level component of the DLM may vary. We illustrate this in Figure 5.1 using the 42.5 km altitude and 35°N to 45°N latitude SOO time series where we show the time series as the blue dots and the background level fit of the model as the orange line for four different

values of σ_{trend}. The values we use for σ_{trend} are $0.0001, 0.0004, 0.01$, and 100. We see that by increasing σ_{trend} the background level component of the DLM is allowed to vary more quickly, until the point where it becomes so large in the case of $\sigma_{trend} = 100$ that the background level component essentially follows the data completely. This last case of $\sigma_{trend} = 100$ is clearly not useful, for the ozone time series or any time series. However, we have found through experience that the odd time the MCMC chain for the DLM can get stuck in this region of σ_{trend}. This is bad since we know for the ozone problem that something around $\sigma_{trend} = 0.0001$ is much more reasonable. To fix this, all we need to do is select a prior distribution which is uniform with a lower bound of 0 and an upper bound of some finite number that we decide σ_{trend} should not be larger than. Again, the lower bound is 0 because a standard deviation parameter cannot be less than 0. As for how to select the upper bound, recall what we are trying to model with this background level component in a stratospheric ozone time series. We are trying to model slow changes in concentration over decadal periods, mainly due to the slowly changing chlorine situation in the stratosphere. In the current state-of-the-art Prais-Winsten Estimation procedure for quantifying these changes, we even assume that the rate of change is constant for periods of time longer than a decade. So, we know that we should not have a quickly changing background level at all in our model. When we present the results of the DLM procedure on the SOO data record in the next chapter, we will choose 0.0004 for this upper bound. This value is used in Figure 5.1 (b), and so this plot gives an idea of the maximum variability of the background level we are allowing for the results we present in the next chapter.

We make one last note about using a uniform prior with an upper and lower bound, which is only related to optimizing the MCMC algorithm. Say that we decide to specify a uniform prior between 0 and 0.0004 for a parameter (like we are) and that for this parameter we use a Gaussian centered at the current state proposal distribution in the Metropolis-Hastings MCMC algorithm (see Section 3.3.1 for the description of this proposal distribution). What we note is that this choice could be equivalently formulated by value constraining the proposal distribution to be between 0 and 0.0004 instead, and having the prior distribution uniform with no bounds (see Section 3.3.2 for a description of value constraining a proposal distribution). The formulated model and MCMC results are essentially the same, the only



(a) $\sigma_{trend} = 0.0001$ **(b)** $\sigma_{trend} = 0.0004$

(c) $\sigma_{trend} = 0.01$ **(d)** $\sigma_{trend} = 100$

Figure 5.1: DLM background level fits with different choice of σ_{trend}. X-axis: Time. Y-axis: Relative Anomaly. Blue: SOO 42.5 km altitude 35° to 45° N latitude data. Orange: DLM background level fit.

difference is that the performance of the MCMC experiment is likely improved. It is improved because the MCMC algorithm will not waste time proposing states that are outside of the prior distribution and will never be accepted into the chain. The only thing that makes this not completely cut and dry is that it may take longer to draw a random value or evaluate a value from a constrained proposal distribution at each step of the Metropolis-Hastings algorithm. When presenting the final results of our model for the SOO data record we will use the option of value constraining the proposal distribution, rather than the prior distribution.

5.2 Sampling the DLM Results and a high-level Overview of the Procedure

We have already hinted that we want to run an MCMC experiment to obtain DLM results for each value of $\boldsymbol{\theta}$ in the MCMC chain and that we want to use all of these results to make inferences about the time series of interest. In this section, we outline the data about the

DLM that we record with each value of $\boldsymbol{\theta}$. This is the last piece of the puzzle for the developed DLM procedure of this thesis work. A high-level overview of the procedure is given at the end of this section.

For each $\boldsymbol{\theta}$ in the MCMC chain, consider that we have ran the DLM filtering algorithm followed by the smoothing algorithm. The result is the smoothed estimates $\mathbf{x}_i^n$ and covariances $\mathbf{P}_i^n$ for $i = 1, ..., n$ for each $\boldsymbol{\theta}$ in the MCMC chain. Now, recall from Section 4.4 that the probability distribution of $\mathbf{X}_i$ given the data $\tilde{y}_n$ is given as, $p(\mathbf{x}_i|\tilde{y}_n) \sim N[\mathbf{x}_i^n, \mathbf{P}_i^n]$. So, what can be done is to take a sample from this Gaussian distribution at each $\boldsymbol{\theta}$ in the MCMC chain, and then concatenate all these samples. This is done for each i, and the result is a sample of the state vector $\mathbf{X}_i$ at each index i that takes uncertainty in $\boldsymbol{\theta}$ into account.

We note that for the probability distribution we have written here as $p(\mathbf{x}_i|\tilde{y}_n)$ from Section 4.4 could be written more explicitly as $p(\mathbf{x}_i|\tilde{y}_n, \boldsymbol{\theta})$ to indicate that when it is calculated we are assuming a given $\boldsymbol{\theta}$ (or a given DLM). Then, with the procedure described in the above paragraph we can consider that we are drawing a sample from the true $p(\mathbf{x}_i|\tilde{y}_n)$ with $\boldsymbol{\theta}$ integrated out.

We have described the procedure for sampling the state vector, now we will consider sampling the DLM model fit. With the discussion of the model fit in Section 4.7, we know that the model fit is given by the probability distribution $N[\mathbf{F}_i\mathbf{x}_i^n, \mathbf{F}_i\mathbf{P}_i^n\mathbf{F}_i^T]$. So likewise, we take a sample from this distribution at each $\boldsymbol{\theta}$ in the MCMC chain for each index i and the result is a sample of the model fit for each index i taking uncertainty in $\boldsymbol{\theta}$ into account. Also, it should be noted that another equivalent method to obtain a sample of the model fit at an index i is, of course, to take the sample of the state vector from $p(\mathbf{x}_i|\tilde{y}_n)$, the way that was described in this section, and then right multiply everything with the matrix $\mathbf{F}_i$.

At this point, we have touched on every aspect of the DLM procedure developed in this thesis. To make the description of the procedure more concrete, we give a high-level overview of its execution in a computer using the diagram in Figure 5.2. The diagram in Figure 5.2 describes the process that takes place each time a new $\boldsymbol{\theta}$ is generated from the MCMC experiment. The entire DLM procedure is just this process repeated for every $\boldsymbol{\theta}$ generated. The yellow blocks in the diagram represent software that performs a specific task, and the blue block represents computer storage space for the samples. The $\boldsymbol{\theta}$ generator block

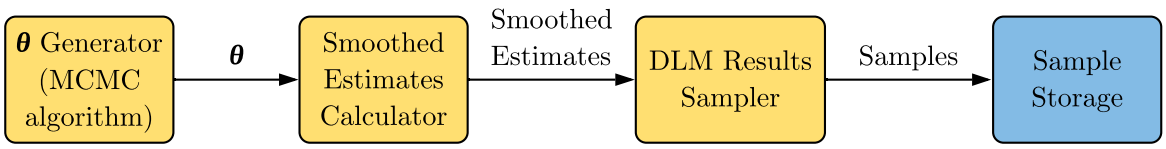


Figure 5.2: High-level Overview of the DLM Procedure

is effectively our Metropolis-Hastings MCMC algorithm, it generates the $\boldsymbol{\theta}$'s. The Smoothed Estimates Calculator then calculates the smoothed DLM estimates given this $\boldsymbol{\theta}$ with the smoothing recurrence relations. If the $\boldsymbol{\theta}$ is the same as it was in the previous iteration (which as we recall from Chapter 3, happens frequently in MCMC experiments), this block should have the previous smoothed estimates stored in memory so that it does not have to redo the calculation in this case. With these smoothed estimates, the DLM Results Sampler then generates the desired samples. For example, a sample of $p(\mathbf{x}_i|\tilde{y}_n)$ from $N[\mathbf{x}_i^n, \mathbf{P}_i^n]$ and a sample of the model fit from $N[\mathbf{F}_i\mathbf{x}_i^n, \mathbf{F}_i\mathbf{P}_i^n\mathbf{F}_i^T]$ at each i, as discussed in this section. The last step is to put the samples generated into computer storage, appending them with the respective samples from $\boldsymbol{\theta}$'s generated previously. Again, this is done until the MCMC experiment is complete, and the result is our desired samples stored in the Sample Storage.

5.2.1 Sampling Specific to Stratospheric Ozone Time Series

Mainly for comparison purposes between the DLM procedure and the Prais-Winsten MLR procedure for stratospheric ozone, there is another sample we can take. Recall that the MLR model we presented estimates the linear trend in ozone over a large time region with the LINEAR PRE and LINEAR POST regressors. To get a comparable data product from the DLM we constructed in Section 4.6.5 we can look at the difference between the background level fit between two time regions. To do this, we take a sample from $N[\mathbf{x}_t^n - \mathbf{x}_s^n, \mathbf{P}_t^n + \mathbf{P}_s^n]$ where t and s are any indices of the time series. We only consider the background level component of this sample because that is all we care about here, and likewise to the explanation in the previous section, we want to take this sample at each $\boldsymbol{\theta}$ in the MCMC chain and concatenate the results to one final sample so that the uncertainty in $\boldsymbol{\theta}$ is taken into account. Referring to the high-level overview of the DLM procedure given in Figure 5.2, this sample can be added

as part of the DLM Results Sampler block. So, here we stress that for whatever types of samples we may be interested in, we can add them to this part of the procedure.

Lastly, the indices t and s that we are interested in are the ones that give us the best comparison to the linear trends estimates with the LINEAR PRE and LINEAR POST regressors. Since these regressors are from the beginning of the data till usually 1997 and then from 1997 to the end of the data, the indices t and s that we are interested in are the ones that span these time regions.

5.3 Fully Specifying the Procedure

In this section, we give all the inputs to the DLM procedure. These are all choices that the modeller must specify for the procedure. These are given below.

1. The DLM model matrices $\mathbf{F}_i, \mathbf{G}_i, \mathbf{W}_i, \mathbf{V}_i$.
2. The DLM priors $\mathbf{x}_0^0$ and $\mathbf{P}_0^0$.
3. What the unknown parameters $\boldsymbol{\theta}$ of the DLM are.
4. The starting point, number of burn-in iterations, number of iterations after the burn-in, and thinning rules (if any) for the Metropolis-Hastings algorithm.
5. The proposal distribution $q(\mathbf{a}|\mathbf{b})$ for the Metropolis-Hastings algorithm.
6. The prior distribution of the DLM unknown parameters $\boldsymbol{\theta}$ for use in the Metropolis-Hastings algorithm.
7. A description of the sampling done with the smoothed DLM estimates at each $\boldsymbol{\theta}$ in the MCMC chain. Discussion of this was had in Section 5.2. Options for this include:
   (a) Sampling the state vector from $N[\mathbf{x}_i^n, \mathbf{P}_i^n]$ at each index i. For this the size of the sample taken at each $\boldsymbol{\theta}$ in the MCMC chain must be specified.
   (b) Sampling the DLM model fit from $N[\mathbf{F}_i\mathbf{x}_i^n, \mathbf{F}_i\mathbf{P}_i^n\mathbf{F}_i^T]$ at each index i. Again, for this the size of the sample taken at each $\boldsymbol{\theta}$ in the MCMC chain must be specified.

(c) Sampling the difference between the background level at two points from $N[\mathbf{x}_t^n - \mathbf{x}_s^n, \mathbf{P}_t^n + \mathbf{P}_s^n]$ (see Section 5.2.1). The size of the sample taken at each $\boldsymbol{\theta}$ in the MCMC chain and the points t and s where the difference is calculated must be specified.

In this chapter, the developed DLM procedure of this thesis work was described. In the next and final chapter before the summary and conclusion, we show the results of the DLM procedure for the example application of stratospheric ozone with the SOO data record.

6 Results

In this chapter, the results of the DLM procedure applied to the SOO monthly mean relative anomaly data record are given. In Section 6.1 the inputs we choose to use for the procedure are specified. In Section 6.2 the results for a single altitude-latitude time series in the SOO data record are shown in detail. Then, in Section 6.3 we show the results for all the SOO altitude-latitude regions. We do this in a way that attempts to convey as much information as possible in heat maps, showing a detailed picture of the historical trends in stratospheric ozone.

6.1 Chosen Inputs to the DLM Procedure

In this section, we outline the inputs to the DLM procedure that we choose by going through all of the items on the list given in Section 5.3. Since the data record we are using is a relative anomaly data record it is already deseasonalized. So, we elect not to use any Fourier form seasonal model components. We use a local level trend model (Section 4.6.4), a multiple regression DLM (Section 4.6.2) with the QBOA, QBOB, SOLAR, ENSO, and AOD regressors, and a first order autoregession DLM (Section 4.6.3). The first three model matrices $\mathbf{F}_i, \mathbf{G}_i$, and $\mathbf{W}_i$ are therefore the same as what was shown in the example in Section 4.6.5 minus the Fourier form seasonal DLMs. We note that the components of the state vector in order are then referred to as the level, trend, 5 regressors, and the AR components. For the last model matrix V_i, we use the square of the standard deviation of the relative anomaly data provided with the SOO data record. For the DLM priors, we use,

$$\mathbf{x}_0^0 = \mathbf{0}, \tag{6.1}$$

and

$$\mathbf{P}_0^0 = \begin{bmatrix} 100 & 0.1 & 100 & 100 & 100 & 100 & 100 & \frac{\sigma_{AR}^2}{1-\rho^2} \end{bmatrix}, \tag{6.2}$$

based on the discussion of Section 4.6.6. So, the unknown parameters $\boldsymbol{\theta}$ for this DLM are

$$\boldsymbol{\theta} = \begin{bmatrix} \sigma_{trend} & \sigma_{AR} & \rho \end{bmatrix}. \tag{6.3}$$

We use a starting point of $\begin{bmatrix} 0.0001 & 0.005 & 0.5 \end{bmatrix}$ for the MCMC with 2000 burn-in iterations, 100000 iterations after the burn-in, and no thinning rules. We choose the proposal distribution to have independent variables such that Equation 3.7 can be used. For each unknown parameter σ_{trend}, σ_{AR}, and ρ we use value constrained Gaussian distributions centered at the current state (see Section 3.3.1 for the description of this proposal distribution) with variances of 0.0001, 0.003, and 0.15 respectively. The value constraints are between 0 and 0.0004, 0 and positive infinity, and 0 and 1 respectively. The prior distribution for $\boldsymbol{\theta}$ is chosen to be uniform over the entire space. Recall from the discussion at the end of Section 5.1.2 that this choice of proposal and prior distributions is equivalent to choosing the proposal distributions to not be value constrained and instead have the prior distribution bounded by the same value constraints. The only reason we do it the way we are instead is because we believe the MCMC fairs better this way.

For reporting our results, we take a sample of size 100 of both the state vector from the distribution $N[\mathbf{x}_i^n, \mathbf{P}_i^n]$ and model fit from the distribution $N[\mathbf{F}_i\mathbf{x}_i^n, \mathbf{F}_i\mathbf{P}_i^n\mathbf{F}_i^T]$ at each $\boldsymbol{\theta}$ in the MCMC chain. Also, we take a sample of size 100 for the difference between the background level at present (2018-02) and 1997-01 and at 1997-01 and the start of the time series (1984-10) at each $\boldsymbol{\theta}$ in the MCMC chain.

6.2 Example Time Series

In this section, we show the results of the DLM procedure in detail for a single time series. The SOO MZM time series for an altitude of 42.5 km and latitude region of 35° to 45° N is shown in Figure 6.1 without any error bars. We have used this time series as an example throughout this thesis, and select it again here since it is in the region where the strong negative trend is seen before 1997 and slightly positive trend is seen afterward. Again, this

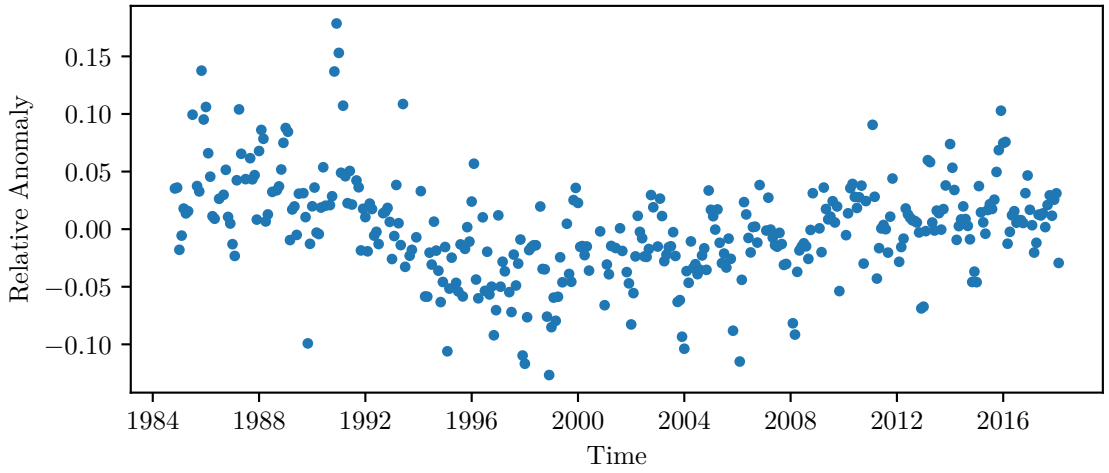


Figure 6.1: SOO MZM Relative Anomaly 42.5 km altitude 35° to 45° N latitude.

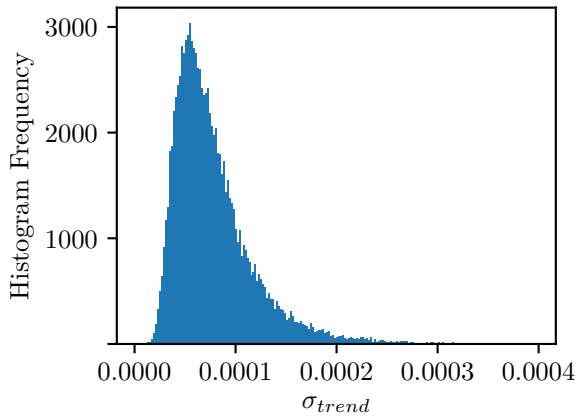
time series and the others in nearby regions were the motivation for choosing the LINEAR PRE and LINEAR POST regressors in the MLR model. In this section, we will see how the DLM procedure we have developed models this data in contrast to the Prais-Winsten estimation, where the results are shown for this time series in Section 2.5.1.

We first show the sampling of $\boldsymbol{\theta}$ from the MCMC experiment by showing the trace plots and histograms for each component of $\boldsymbol{\theta}$ separately in Figure 6.2 (see Section 3.2.2 and Section 3.4.2 for the description and discussion of these MCMC visual diagnoses). Before looking at any other results we always want to look at these to make sure that the MCMC chain has converged. If it has not then all the other results would be less accurate. Furthermore, if the MCMC experiment was set up very badly then the other results could even be unintelligible. But, from the trace plots shown here, we see that the MCMC chain has converged fairly well. Recall that in Chapter 3 of this thesis on MCMC, all we have given as a means to assess convergence is looking at these visual diagnosis illustrations, as this is what is often done in practice. There is no great way to assess convergence in MCMC, but there do exist more rigorous methods. One way we know of involves running multiple MCMC chains. Better assessing the convergence of these MCMC experiments should probably be something that is looked into as future work. But, recall that when the chain converges these histograms become the shape of the marginal distributions of $p(\boldsymbol{\theta}|\tilde{y})$. We are confident that we see close

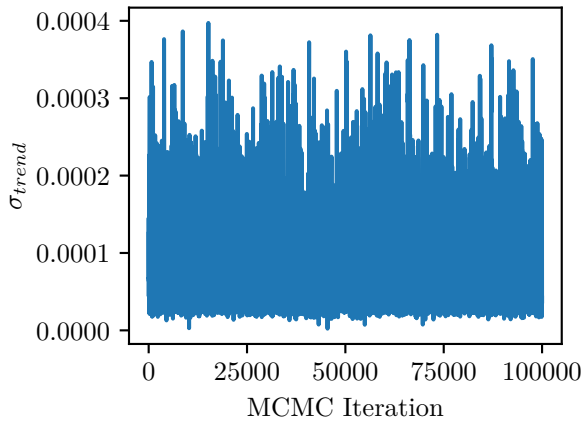
to these marginal distributions here.

To give an idea on how the varying $\boldsymbol{\theta}$ changes the result of the DLM we show the DLM model fit for a few different $\boldsymbol{\theta}$'s in the MCMC chain in Figure 6.3. With this, we get a sense of how changing the DLM input parameters $\boldsymbol{\theta}$ affects the results of the model. But, note that we do not choose any $\boldsymbol{\theta}$ here, these are all $\boldsymbol{\theta}$'s that were obtained from the MCMC experiment, so they are all reasonably probable $\boldsymbol{\theta}$'s for the given data. Similarly, we also show the background level component of the model, which we call the background level fit, for these same $\boldsymbol{\theta}$'s in Figure 6.4. We see in these two figures that the fits are all fairly similar for different $\boldsymbol{\theta}$. This is generally the case for all altitude-latitude regions of the SOO data record. So, it is not surprising that we find that the errors from the DLM itself are much larger than the errors resulting from the uncertainty in its input parameters $\boldsymbol{\theta}$. This means that for most altitude-latitude regions the error estimations for a DLM with an optimal $\boldsymbol{\theta}$ are close to the error estimation from the whole DLM procedure. It will only be a slight underestimate and for some altitude-latitude regions even unnoticeable. Of course, we still need to do our due diligence and account for these errors, with it mattering for certain altitude-latitude regions and for the other data records (for ozone or other species) that the DLM procedure will be applied to.

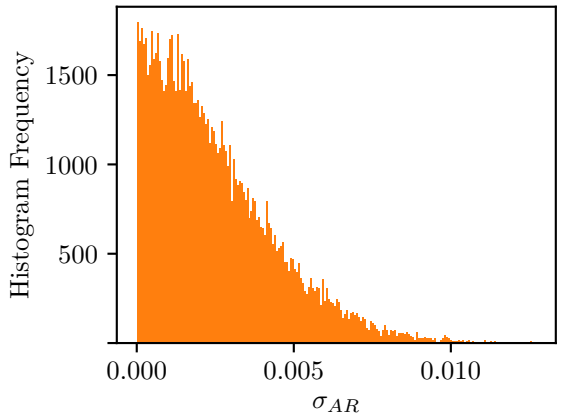
Next, we show the results of sampling the state vector at a given index i from $N[\mathbf{x}_i^n, \mathbf{P}_i^n]$ as described in Section 5.2. We can say that this is a sample from the random vector $\mathbf{X}_i$. In Figure 6.5 we show the sample of $\mathbf{X}_{50}$ (index 50 corresponds to December 1988 in this time series) by generating histograms for each of its components. We put a red line at the mean of these and an orange line at the 5th and 95th percentiles. Note that they all appear to be very close to Gaussian distributed. This is because for these, as described in Section 5.2, we are sampling from a Gaussian distribution at each $\boldsymbol{\theta}$ of the MCMC chain and concatenating the results. So, what we are seeing here is related to how the DLM hardly changes for different $\boldsymbol{\theta}$'s of the MCMC chain, which was already discussed and shown in Figures 6.3 and 6.4. Now, we should understand that of course if $\boldsymbol{\theta}$ had more of an effect we would potentially see samples that do not appear to be Gaussian. We can see now from this argument what was meant in the last paragraph when we said that the DLM itself has much larger error than errors resulting from the uncertainty of the input parameters $\boldsymbol{\theta}$. It is because the spreads



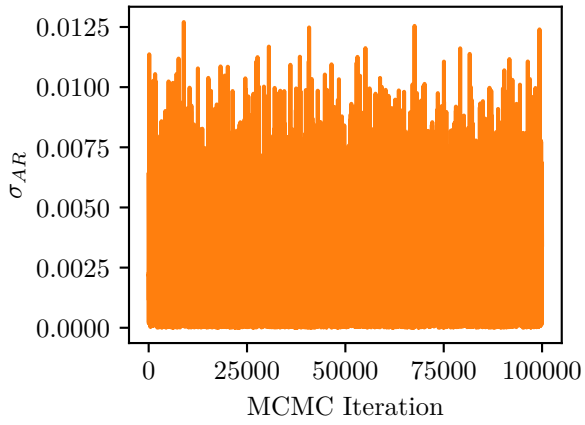
(a) σ_{trend} Histogram



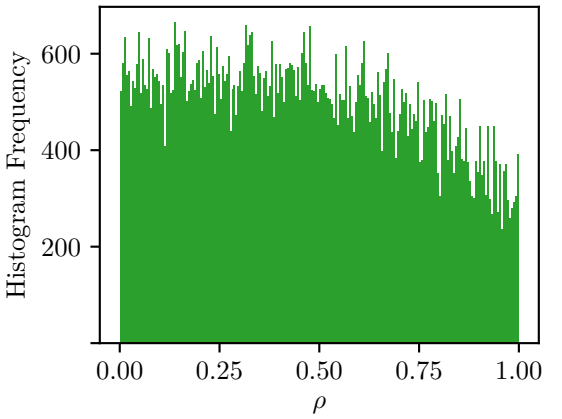
(b) σ_{trend} Trace Plot



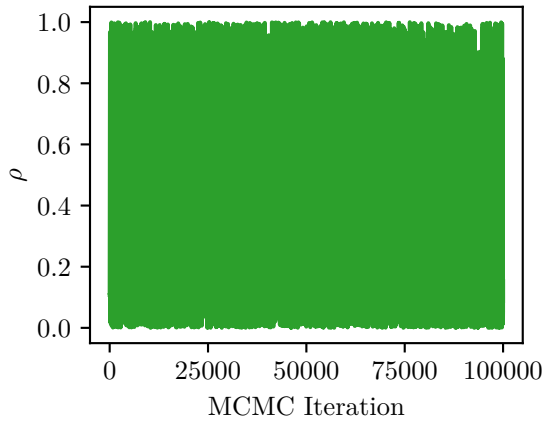
(c) σ_{AR} Histogram



(d) σ_{AR} Trace Plot



(e) ρ Histogram



(f) ρ Trace Plot

Figure 6.2: MCMC Histograms and Trace Plots

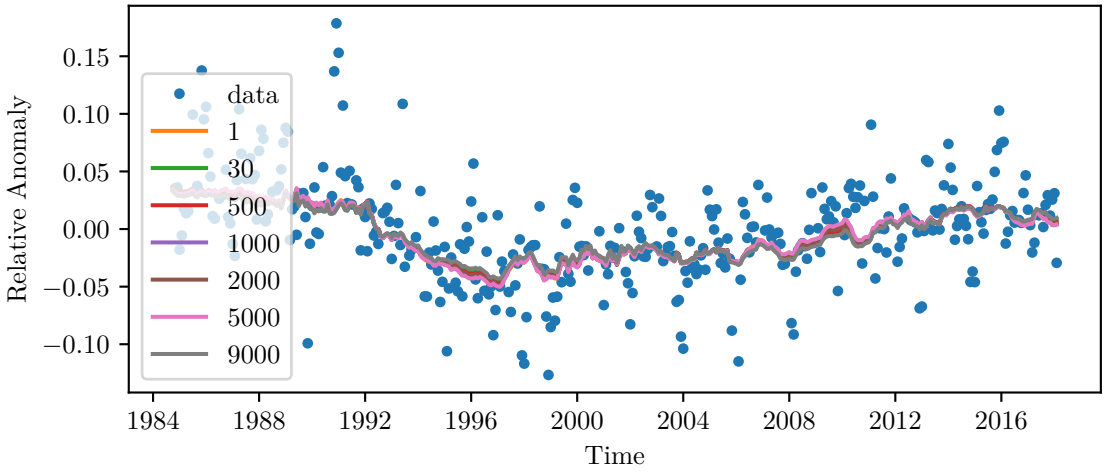


Figure 6.3: DLM model fits at various $\boldsymbol{\theta}$'s in the MCMC chain. Numbers 1, 30, 500, etc. represent the index of the MCMC chain.

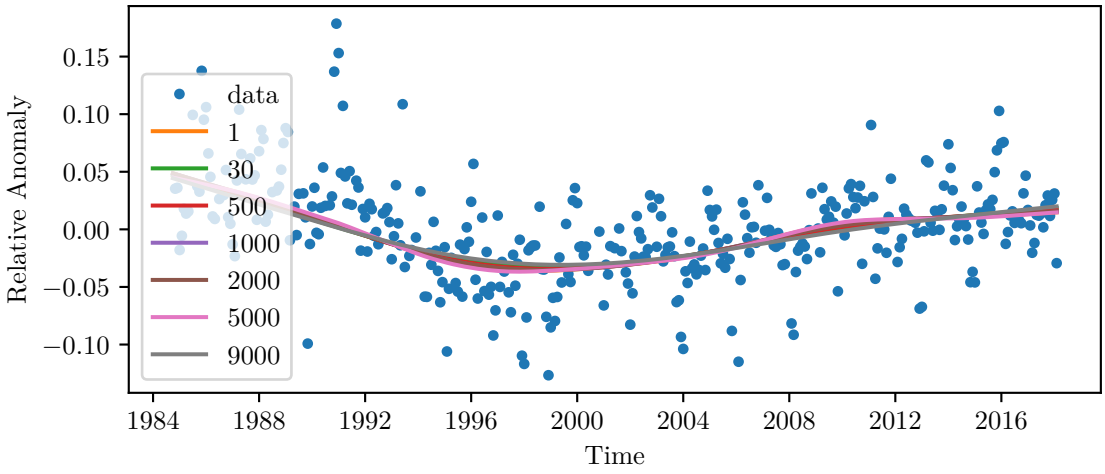


Figure 6.4: DLM background level fit at various $\boldsymbol{\theta}$'s in the MCMC chain. Numbers 1, 30, 500, etc. represent the index of the MCMC chain.

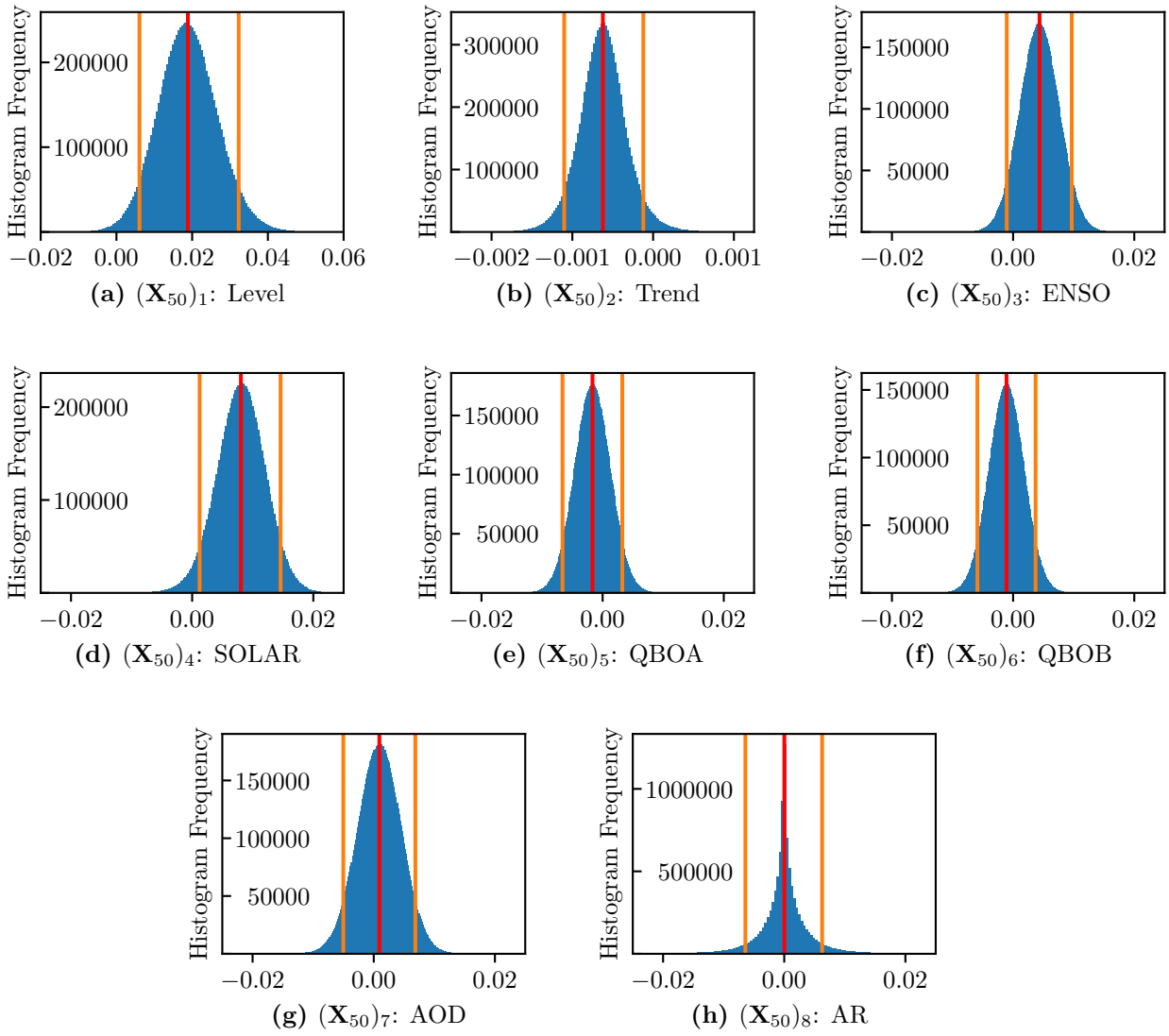


Figure 6.5: A sample of the state vector at index 50 (December 1988).

of these samples are not increased much by varying $\boldsymbol{\theta}$, instead, the spreads are basically the same as what results from the Gaussian that describes the error of the DLM.

We also generate a histogram from the sample of the DLM model fit at index 50. This is shown in Figure 6.6. To obtain this sample, we simply right multiply with the matrix $\mathbf{F}_{50}$ to each sample of $\mathbf{X}_{50}$ (recall from Section 5.2 that this is the same as drawing a sample from the model fit Gaussian distribution $N[\mathbf{F}_{50}\mathbf{x}_{50}^n, \mathbf{F}_{50}\mathbf{P}_{50}^n\mathbf{F}_{50}^T]$).

Recall that for our chosen DLM the regression coefficients of the multiple regression DLM component do not vary with index i. So, for these elements of the state vector, we

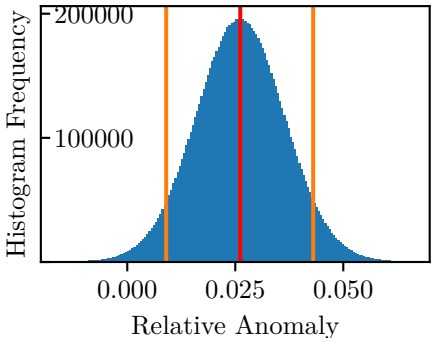


Figure 6.6: A sample of the model fit at index 50 (December 1988).

can concatenate the samples at each i together to get a single larger sample, opposed to just considering them at index 50 like was illustrated in Figure 6.5. The histograms for these concatenated samples for the five regressors are shown in Figure 6.7.

We can, in a sense, visualize all $i = 1, ..., n$ model fit samples rather than just single ones like we have in Figure 6.6. We show this in Figure 6.8 where the mean value of the samples at each index is shown as the orange line and the 5th and 95th percentiles are the ends of the shaded region. Similarly, for the background level fit we show the sample means and 5th and 95th percentiles in Figure 6.9.

Lastly, we show the histograms resulting from the samples that we chose to create for the difference between the background level fit from 2018-02 to 1997-01 and from 1997-01 to 1984-10 in Figure 6.10. These illustrations show an important result for stratospheric ozone. We multiply the samples by 100% multiplied by 120 total months in 10 years divided by the number of months between where the differences were calculated (i.e. $t - s$ referring to the t and s in Section 5.2.1) so that the units are converted to % change per decade, the commonly used unit is stratospheric ozone trend literature. The means of the samples show, as we already know, that before 1997 there appears to be a decrease in ozone and after 1997 a smaller increase. But more importantly, the Gaussian looking samples hardly cross zero in both cases, showing us that there is a great deal of certainty that ozone did indeed decrease and then increase. We see from the orange lines that well over 95% of the sample is negative for the before 1997 case and well over 95% is positive in the after 1997. In fact, the percentile of the zero point for the before case is 99.96% and for the after case 0.15%.

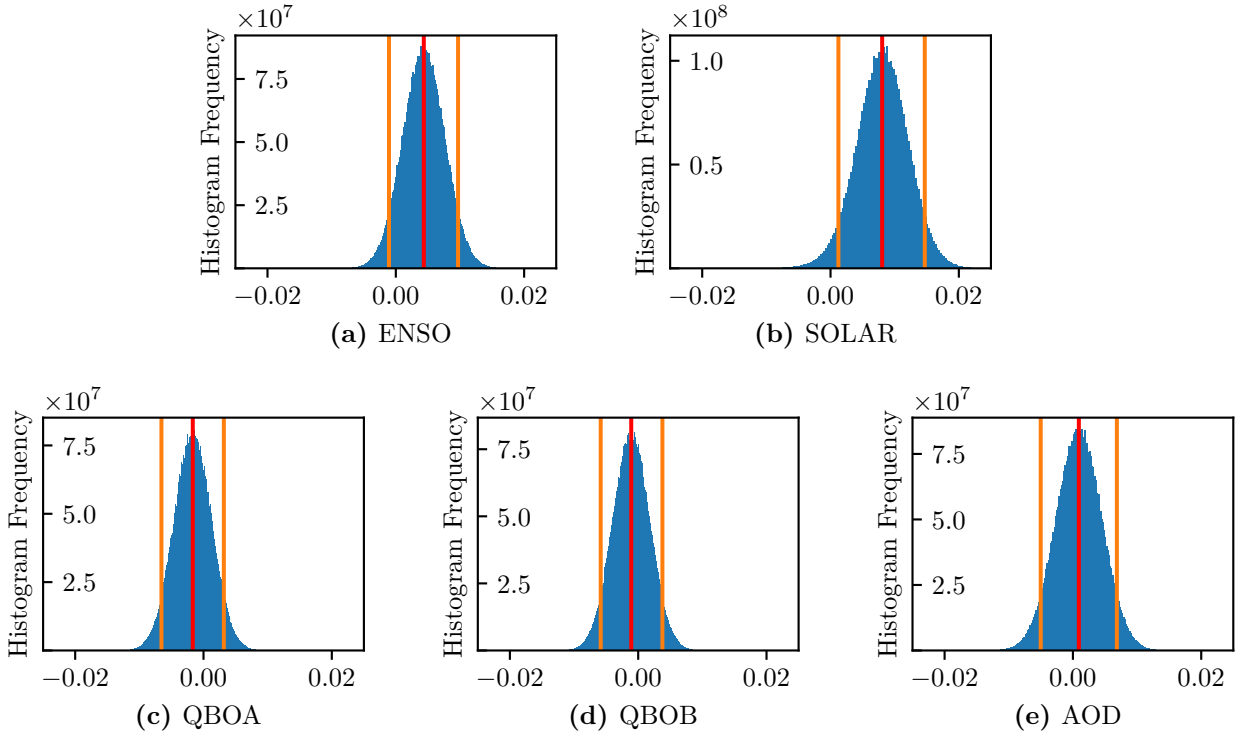


Figure 6.7: A sample of the regression coefficient components of the state vector where the samples from all indices i are concatenated together.

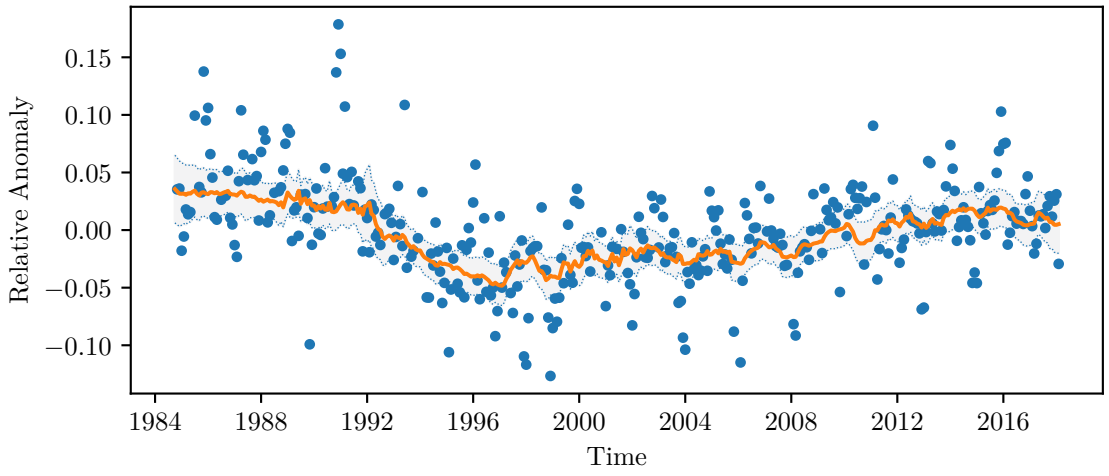


Figure 6.8: Mean model fit with 5th and 95th percentiles.

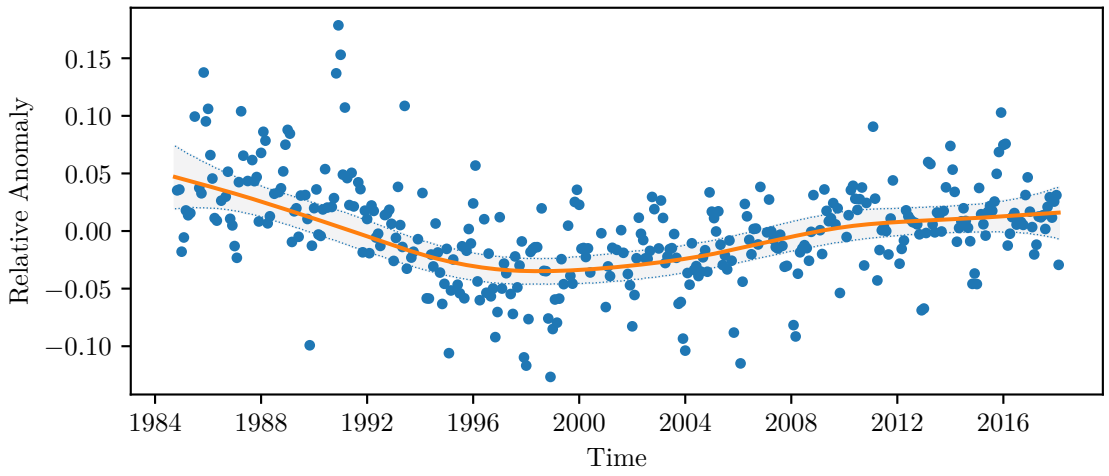
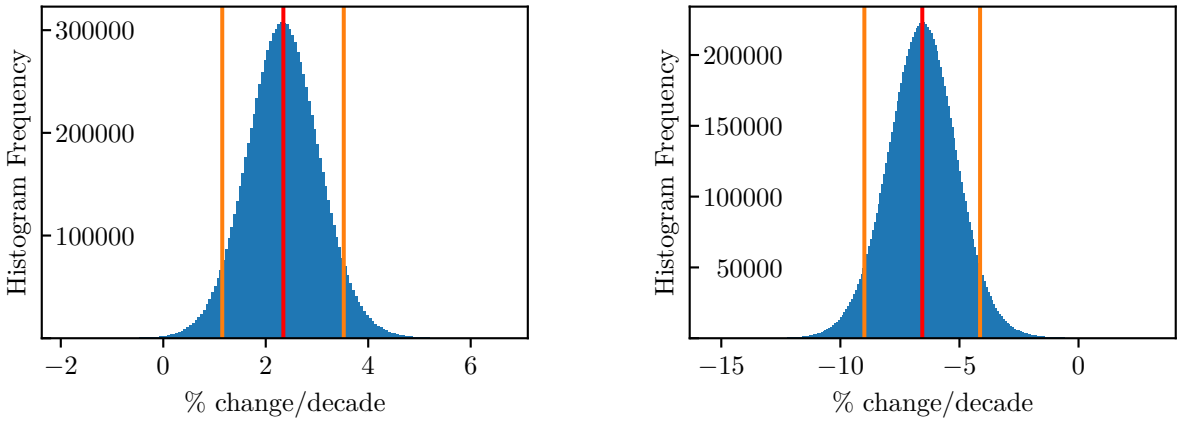


Figure 6.9: Mean background level fit with 5th and 95th percentiles.



(a) 2018-02 to 1997-01

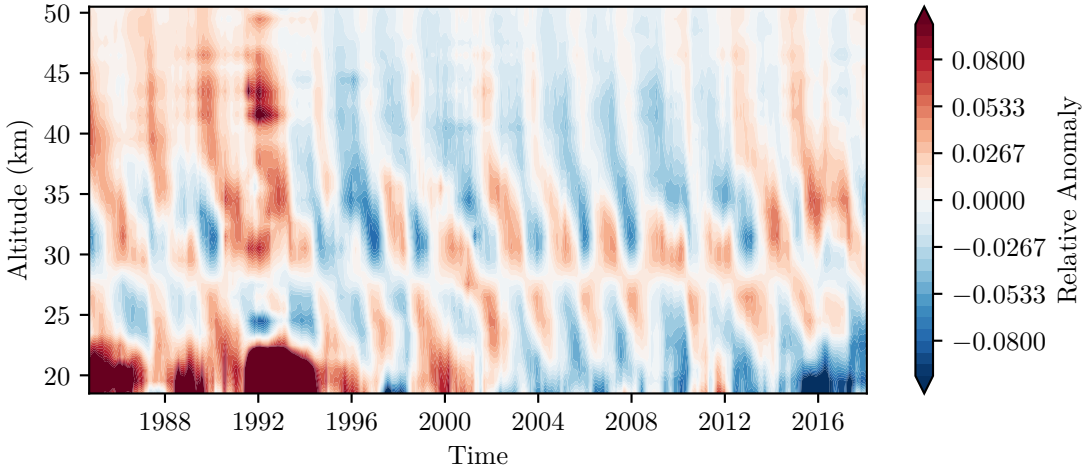
(b) 1997-01 to 1984-10

Figure 6.10: A sample of the background level differences.

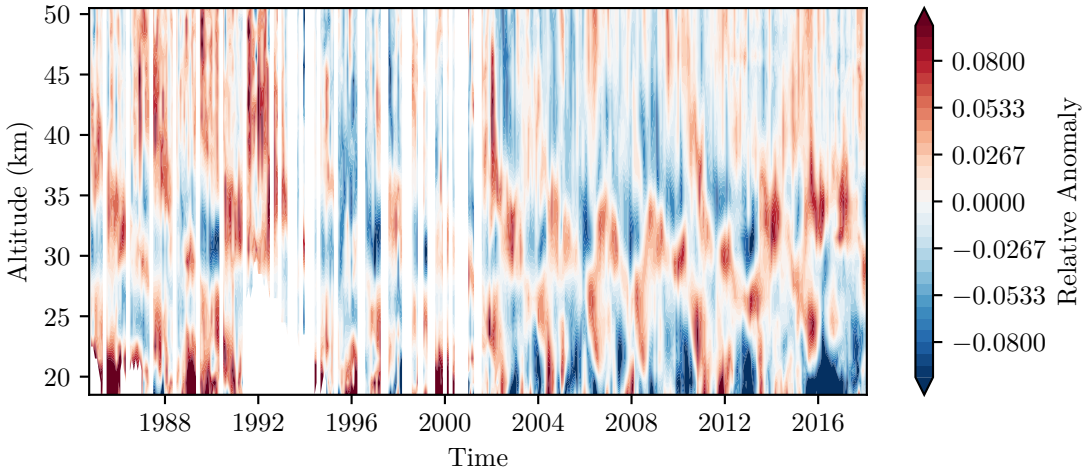
6.3 Results for the SAGE II/OSIRIS/OMPS Data Record

In this section, we convey as detailed as possible all the results shown in detail in the last section for a single time series, but for all altitude-latitude regions in the SOO data record. Again, the first thing we want to do is make sure that the MCMC chains for each altitude-latitude region have converged. We show all these trace plots and histograms in Appendix V, where further discussion about the results of these is also had. All of the chains show that they are fairly well converged. Now, consider trying to show the model fit sample illustrated in Figure 6.6 as a histogram, but for all indexes of the data and for all altitude-latitude regions. One way we can present this is to just show the means of these samples in a set of heat maps. For instance, we can make an individual heat map of these means for each latitude region where the y-axis covers all the altitudes and the x-axis all the time. We show this in Figure 6.11 for the example latitude region of 5° to 15° N. The rest of the latitudes are shown in Figures 6.12 and 6.13. With these illustrations, we see the point estimates of the DLM model fit for the entire SOO data record. In Figure 6.11 we also show the raw SOO data that is being fit too.

Now, we can make similar illustrations for any component of the DLM state vector. So, this brings us to what is thus far the primary application for the DLM procedure developed in this thesis, to quantify stratospheric ozone trends from the SOO data record. Again, consider the trend sample illustrated in Figure 6.5 (b) as a histogram. We show the means of these trend samples for the example latitude region of 5° to 15° N in Figure 6.14. Recall from Section 4.6.5 that with the way we have selected our DLM, this component of smoothed estimates of the state vector is equivalent to the derivative of the background level fit. We have also multiplied this quantity by 100 x 12 months x 10 years so that the units of it become percent change per decade. Since this unit is commonly used in stratospheric ozone trend literature. So, this illustration combined with the ones for the other latitude regions gives a full picture of the trends estimated by the chosen DLM. A feature of this illustration that was not in the illustrations for the model fit is that we attempt to give an idea of



(a) Mean model fit



(b) Raw SOO data

Figure 6.11: Mean model fit for latitude 5° to 15° N.

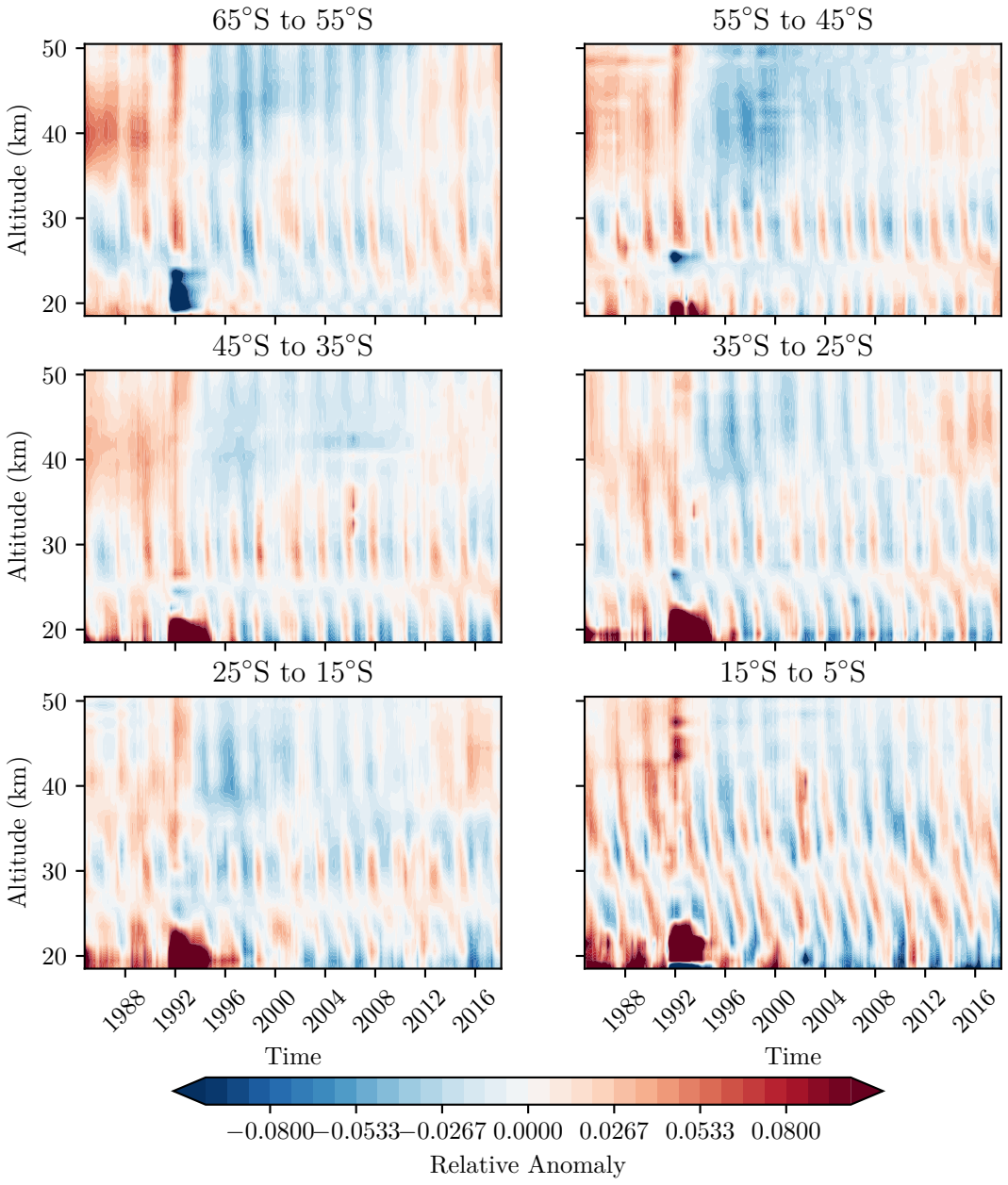


Figure 6.12: Mean Model Fits.

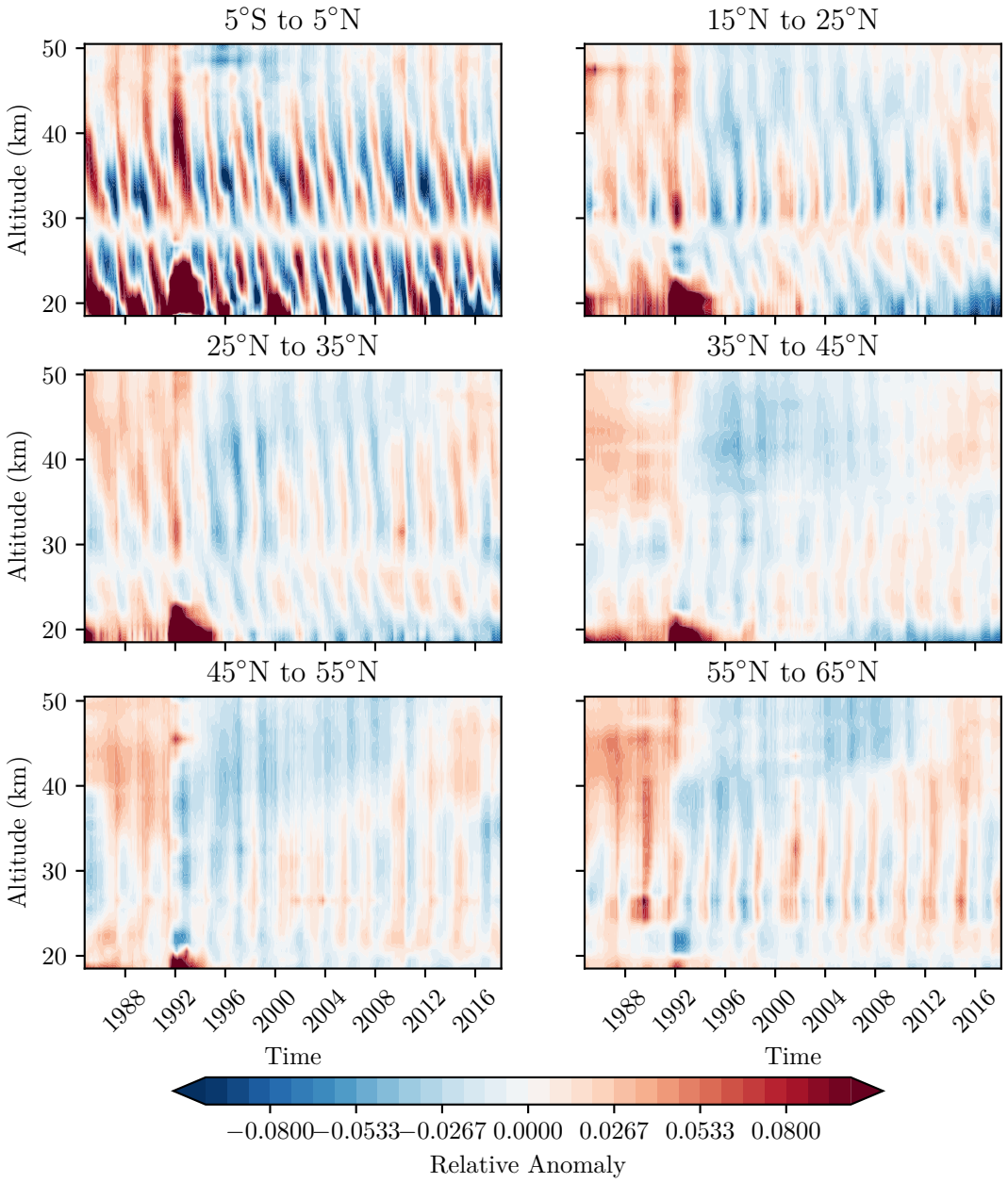


Figure 6.13: Mean Model Fits.

the certainty of the sign from the sample. To do this we overlay solid, dashed, and dotted grey contours on the heat map that indicate what percentage of negative (or positive) values that the sample contains. The solid dashed and dotted lines represent 80%, 90%, and 95% respectively. Whether a given dotted contour, for example, is showing the bound for 95% positive values or 95% negative values is obvious from the surrounding color. When the surrounding color is positive then the contour is showing the 95% positive values bound and when it is negative it is showing the 95% negative values bound. We have used these dotted, dashed, and solid contours before in reporting the Prais-Winsten trend results in Figure 2.12. The purpose of these is the same here as it was there, we are trying to indicate the percent confidence in the sign of the point estimate in the trend.

The illustrations for the rest of the latitude regions are shown in Figures 6.15 and 6.16. The combinations of all of these give a very complete picture of the trends in stratospheric ozone concentration in the given latitude regions. There is one important conclusion that we can draw from these. It is already known that ozone concentrations in the middle to high latitudes of the upper stratosphere decreased from 1984 until about 1997 and increased to a lesser degree after. We can actually see this clearly now in these illustrations for the middle to high latitudes, where in the upper altitudes we see, for the most part, a border between the blue and red colors around the year 1997. This is probably most pronounced in the 55° to 45°S illustration. The important conclusion that we can draw however is that, for the most part, in these regions, the increases after 1997 were stronger closer to 1997 and started to die off afterward (i.e. as time has gone on the increase has lessened). This has been suspected for some time now by the atmospheric remote sensing community, but to our knowledge, never quantified or presented until now.

For the regression terms, we want to make similar illustrations. Conveniently, we lose the time dimension because we can concatenate the samples across time like in Figure 6.7. So, with samples like the ones shown in Figure 6.7 for all altitude-latitude regions, we can illustrate the means for each regressor on a single heat map with dimensions of altitude and latitude. These are shown in Figure 6.17. We also use the dotted dashed and solid contours here for an indication of the certainty in the sign of the point estimate. These illustrations are directly comparable to the analogous illustrations generated from the Prais-

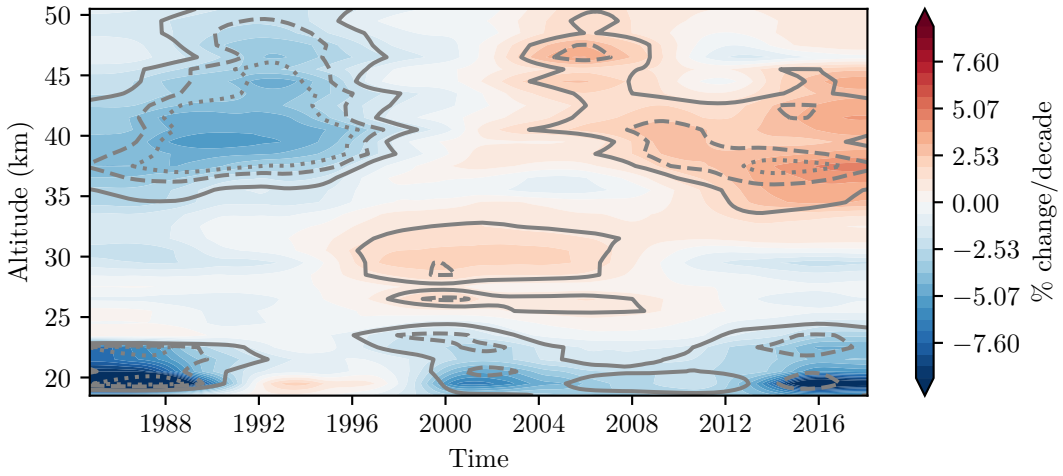


Figure 6.14: Derivative of background level fit for latitude 5° to 15° N, converted to units of percent change per decade.

Winsten in Figure 2.11. We use the same colormap in both figures for comparison purposes. The conclusion of the comparison is that they agree quite well. The astute reader may have noticed that for the example time series of 42.5 km altitude and 35° to 45° N latitude we have used commonly throughout this thesis, that the Prais-Winsten procedure estimates a fairly large QBOB term, whereas this QBOB term is quite minimal in the DLM procedure's estimate (refer to Sections 2.5.1 and 6.2). This is an interesting difference for this region, but when looking at the entire picture for the SOO data record with these heat map illustrations, we see that the QBOB terms agree fairly well in general. But again, the difference is still there is this region.

For the background level differences, we have a similar scenario to the regression terms. With samples like the ones shown in Figure 6.10 for all altitude-latitude regions, we can again illustrate their means on a single heat map. They are shown in Figure 6.18 with the dotted, dashed, and solid contours. These statistics were mainly calculated in the first place for comparison to the results from the trend estimations from the Prais-Winsten procedure. We can now directly compare these to the Prais-Winsten illustrations in Figure 2.12. We re-show these here in Figure 6.19 for comparison. We see from Figure 6.18 that the structures are very much similar to the Prais-Winsten results illustrations. The magnitudes of the

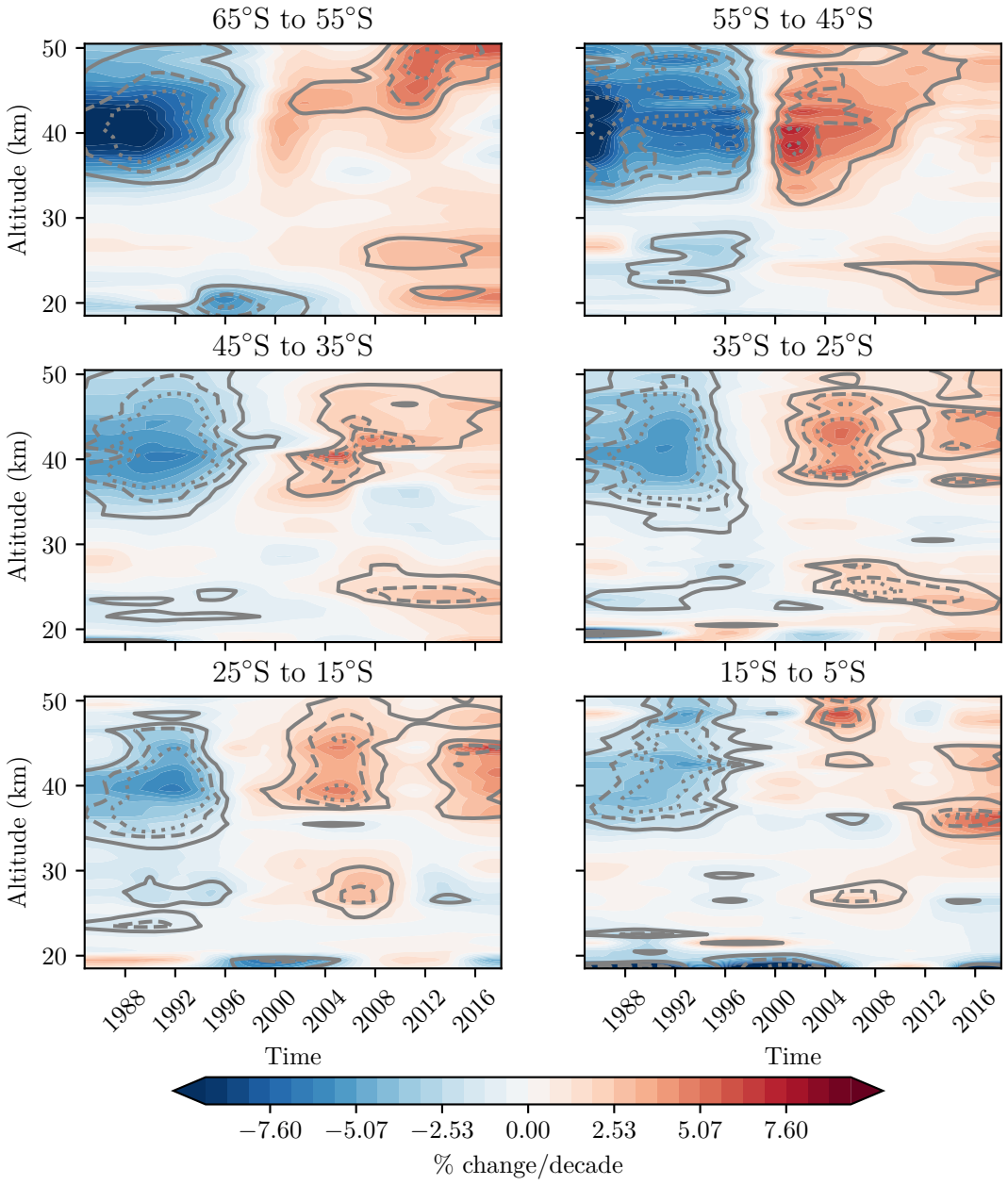


Figure 6.15: Derivative of background level fits, converted to units of percent change per decade.

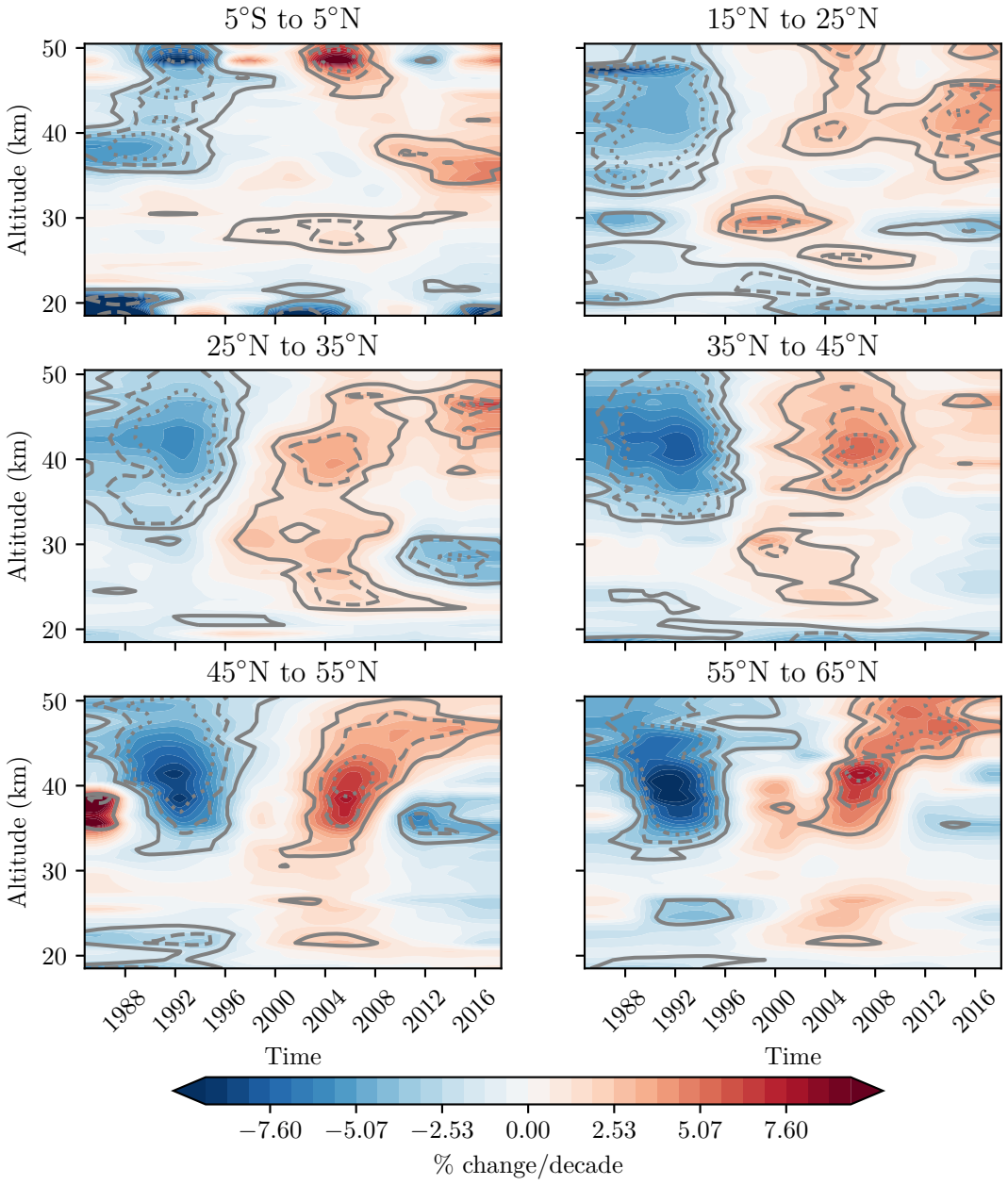


Figure 6.16: Derivative of background level fits, converted to units of percent change per decade.

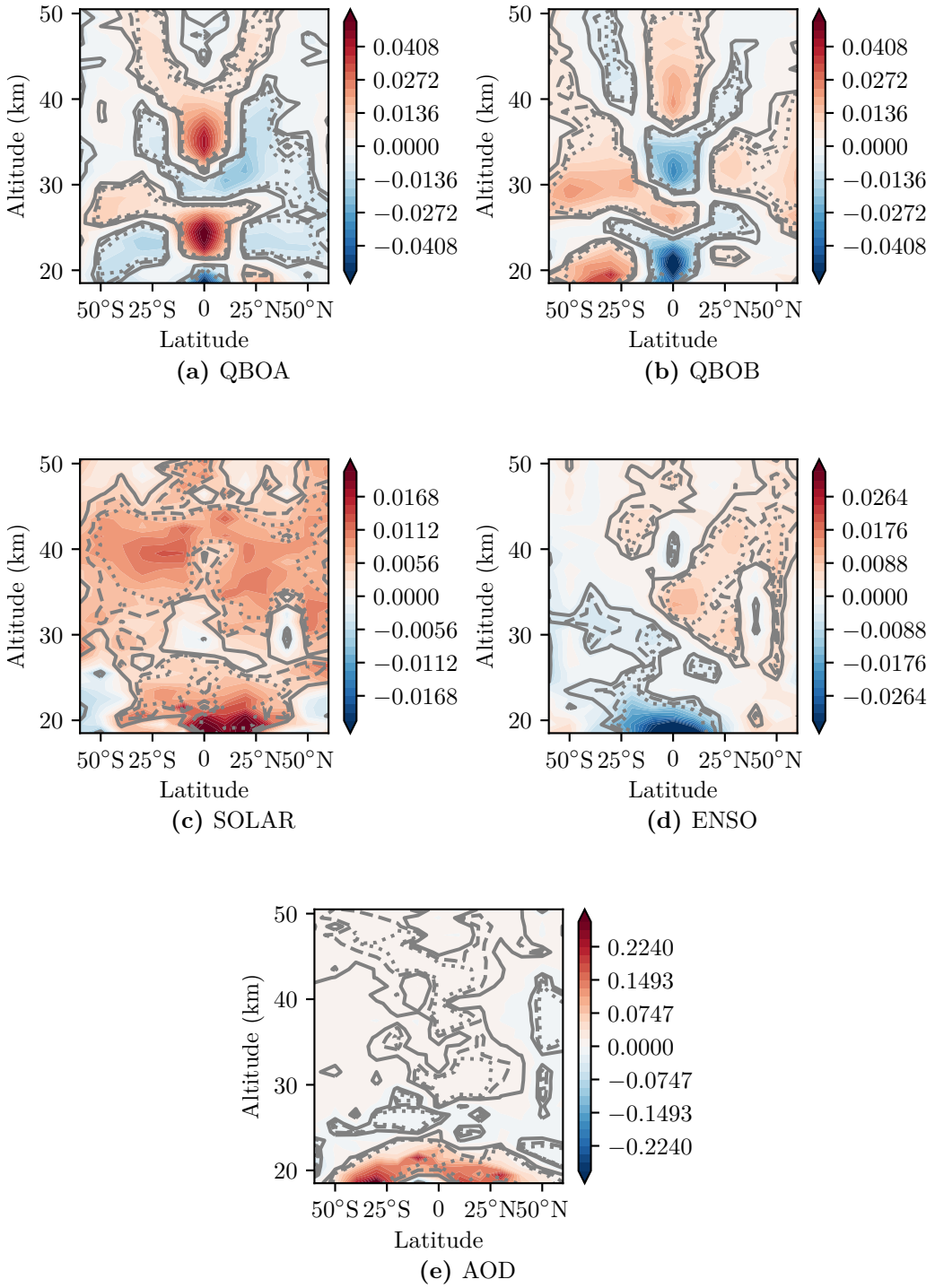
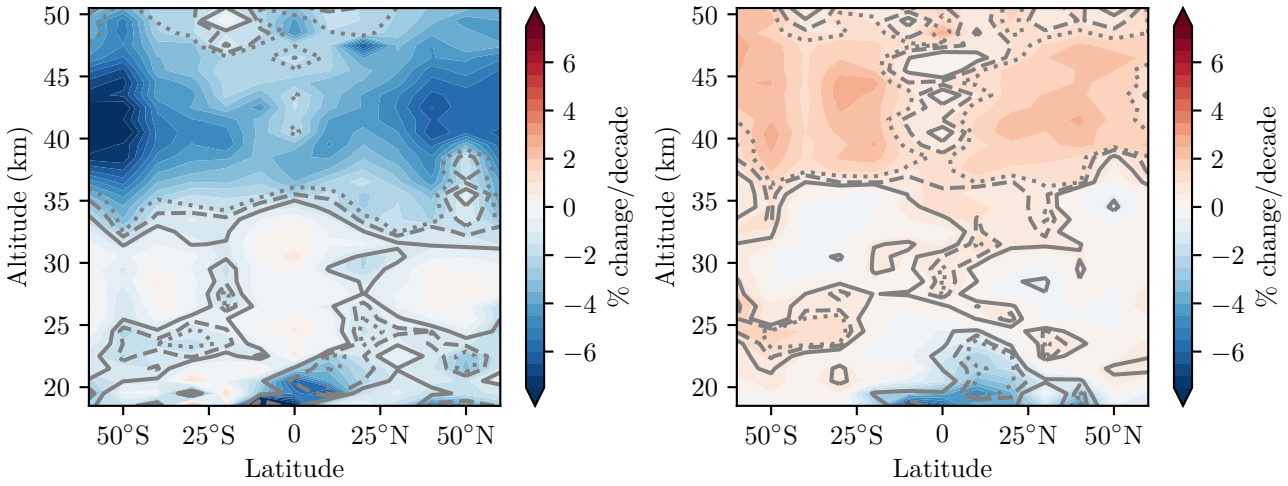


Figure 6.17: Estimated Ozone Influencing Phenomena Coefficients.

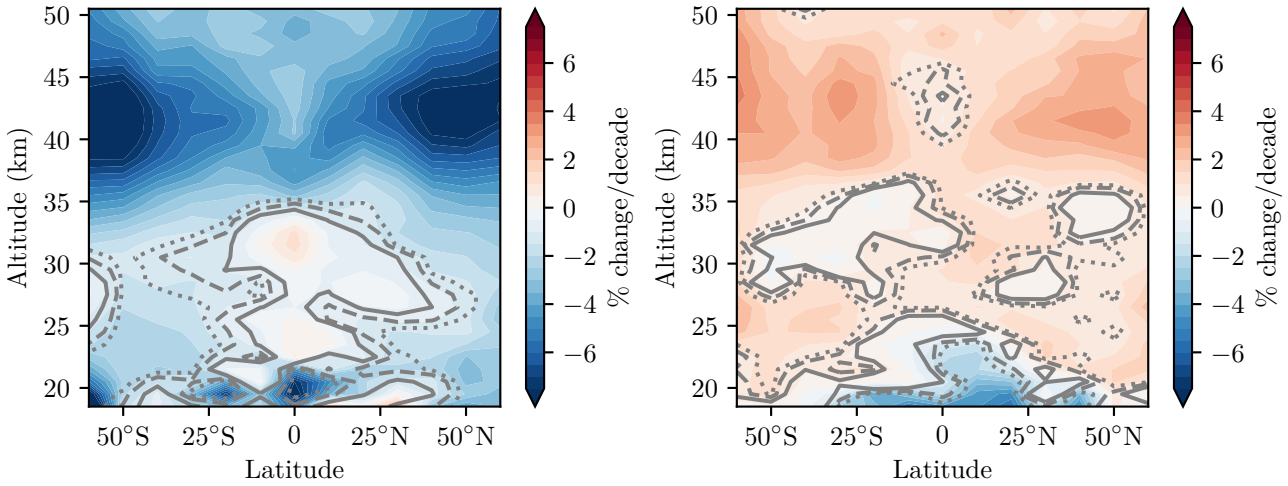
Prais-Winsten illustrations are slightly larger in many regions. This typically works out this way when the inflection point for the LINEAR PRE and LINEAR POST regressors for the MLR model is not well suited for the time series. Also, the confidence levels, as shown by the dotted, dashed, and solid contours, are in general lesser in Figure 6.18 than in Figure 6.19. This could be because the magnitudes are smaller in Figure 6.18 or because the DLM procedure takes uncertainty in ρ into account when estimating errors while the Prais-Winsten procedure does not.

In this chapter, we showed the results of the DLM procedure in detail for a single time series in the SOO data record and for the entire SOO data record. We started by clearly defining our chosen input parameters to the DLM procedure so that the work can be replicated if desired. By showing the results in detail for a single time series, we gave a sense of what the data from the DLM procedure looks like at the lowest level and how to interpret it. Then with the entire SOO data record, we chose to only illustrate the important parts and tried to illustrate them intelligently. With this, we painted a very complete picture of the trends in ozone concentration in the stratosphere from 1984 to present (2019) using the SOO data record and made some noteworthy conclusions.



(a) 1997-01 to 1984-10 **(b)** 2018-02 to 1997-01

Figure 6.18: Background level differences, converted to units of percent change per decade.



(a) LINEAR PRE **(b)** LINEAR POST

Figure 6.19: LINEAR PRE and LINEAR POST regression coefficients, converted to units of percent change per decade.

7 Summary and Conclusion

We conclude that the DLM procedure developed in this thesis is, in general, better to use for quantifying trends in the SOO data record than the Prais-Winsten MLR procedure. There are two reasons we make this claim. One is that the shape of the LINEAR PRE and LINEAR POST regressors in the MLR model is not well suited for many SOO altitude-latitude regions, and the other is that the Prais-Winsten procedure does not report its errors considering uncertainty in the autocorrelation parameter ρ, while the DLM procedure does. We discuss these two points further in the next two paragraphs.

The LINEAR PRE and LINEAR POST regressors constrain the shape of the trends for the MLR model. So, this can be thought of as external information that is imposed upon the model by the modeller. In this light, an argument for the DLM procedure could be that instead of imposing a shape, the ozone data alone should inform us about the shape of the trend. This is exactly what the DLM procedure does. However, to be fairer in this argument between MLR and the DLM procedure, it is not unreasonable in statistics to make big constraints on models if the constraint is known to be true or if some specific goal is desired. It is sometimes even necessary to make constraints so that reasonable results can be obtained. But, for many altitude-latitude regions of the SOO data record, it is difficult to select a set of usable predictors for the MLR model that adequately represents the trend in the time series. Still, the MLR model can be useful in obtaining what it was made for, the linear ozone trend estimation over a wide time region. And, it is of course fairly well suited for the SOO time series in the middle to high latitude upper altitude regions, that were really what was had in mind when the LINEAR PRE and LINEARE POST regressors were constructed. But even still, with the example shown in this thesis with the 42.5 km and 35° to 45° N SOO time series, we see that the DLM background level fit is probably also better suited in this region as well. Furthermore, the different time series in these regions

have different "inflection points" (as we see from the DLM procedure's results given in this thesis, and if not, from the raw data). Perhaps if the MLR model is still going to be used, a procedure could be developed to find some sort of optimal inflection point of the LINEAR PRE and LINEAR POST regressors for each altitude-latitude region time series in the SOO data record. For some of them, this will be tricky however, because the LINEAR PRE and LINEAR POST regressors do not suit the data at all. Lastly, we note that the LOTUS initiative is beginning to recognize this advantage of DLMs over MLR models when they say "being that the DLM does not constrain the shape of the trend terms, it may ultimately make it better suited for ozone trend analysis" (Petropavlovskikh et al., 2019).

The DLM procedure accounts for uncertainty in the autocorrelation parameter ρ, while the Prais-Winsten procedure does not. In the DLM procedure, ρ is a parameter in $\boldsymbol{\theta}$, therefore its uncertainty is accounted for. Something further to note concerning this is that it was shown in this thesis (Figures V.9 through V.12 in Appendix V) that for the majority of altitude-latitude regions of the SOO data record, the MCMC experiment indicates that a wide range of values for this parameter are possible (i.e. the data is not strongly indicating a particular value of ρ is highly probable). Again, the DLM procedure theoretically accounts for the uncertainty associated with these many values of ρ being possible, and the Prais-Winsten estimation does not. With our understanding of how the Prais-Winsten procedure works, we know that ultimately the procedure finds an optimal ρ, but then the results that it reports are simply the GLS estimation results assuming that this is the correct parameter ρ which represents the data with no error. So, the DLM procedure gives a more accurate estimation of errors in this regard, and there is not a great way to amend this for the Prais-Winsten estimation (although something could probably be devised if desired. Something that likely includes Monte Carlo integration's).

To summarize this thesis, we started with a background chapter that contained material about what was in place at the outset of this thesis work. This included the generation of the SOO data record from the three respective smaller data records and the Prais-Winsten MLR procedure that is commonly used to estimate trends in ozone. In this background chapter, we provided some theoretical details behind the Prais-Winsten estimation procedure, which should make this information easier to access for somebody who is not a statistician. The

results of this procedure in modelling and predicting trends with the SOO data record were then shown. In the chapter that followed, we covered the topic of MCMC. This was done independently from the rest of this thesis, and served as an introductory understanding into the topic, with it being such an essential part of the DLM procedure.

The next chapter was on the DLM. We started by giving two introductory DLMs called the multiple regression DLM and the local level DLM. These served as motivation for the DLM and allowed us to see that it is useful to specify DLMs in a general form. After this general form DLM was given, we gave the recursive estimation equations used to estimate its parameters. This was done before a significant portion of this thesis was spent on the theoretical background of these equations, where several different statistical criteria leading to them were shown. This part of the thesis should serve as a good reference to this material, because it is difficult to tread through all of these statistical derivations with multiple other sources that use different notations and have fewer details. So likewise to the Prais-Winsten procedure, we have made this material easier to access for the reader. At the end of this chapter, we showed how to construct specific DLMs for modelling time series data.

In the next chapter, we described the DLM procedure of this thesis work. We started by explaining how to use MCMC experiments to obtain a sample from the probability distribution of a DLM's unknown input parameters. We then described how running the DLM estimation and sampling from its results for each chain in the MCMC experiment is effectively the DLM procedure developed in this thesis. In the next chapter, as an example application for the DLM procedure, we showed its results for the SOO ozone data record. With this, we quantified trends in middle and tropical latitude stratospheric ozone with a procedure similar to procedures that are just starting to be adopted for this problem (Ball et al., 2017; Ball et al., 2018; Laine et al., 2014), and have painted a clear and detailed picture of the quantified trends with Figures 6.14, 6.15, and 6.16.

7.1 Suggestions for Improvement and Future Work

There is a lot of potential improvements one could think about related to the MCMC algorithm. In the DLM procedure, we use a simple Metropolis-Hastings algorithm. For the ozone

problem, some have used slightly more advanced algorithms (Laine et al., 2014). Adaptive metropolis algorithms and Hamiltonian Monte Carlo algorithms are a few MCMC algorithms that could be looked into for improvements. However, we stress that improving the MCMC algorithm is not really going to improve the results of the developed procedure, unless we are wrong in our assessment in this thesis, for the ozone problem for example, that the MCMC experiments are pretty much converged. We recall the discussion that was had in Chapter 3 that a better tuned MCMC algorithm only improves the rate of convergence of the MCMC experiment. So, if an MCMC algorithm is developed that converges 1.5 times faster than a previous one, then the same result could have been achieved by running the previous algorithm for 1.5 times longer. So, when trying to get slight improvements with more advanced algorithms this is just something to keep in mind.

Still, new algorithms could be implemented into the software that was developed for this thesis work. Open-source MCMC software, which of course use more advanced MCMC algorithms, could also be looked into. But, from a glance at them, we got the impression that they do not give as much freedom as we desire for this DLM application. Certain things about the complete algorithm of the DLM procedure can be optimized (i.e. the smoothing recurrence relations do not need to run each MCMC iteration), which we are not sure can be done with open-source MCMC software. Others in the community have used open-source software (Alsing, 2019). But, we find this software to be slower in convergence time than the software we have developed when using a decently tuned proposal distribution for reasons that are unknown to us. Using open-source software or implementing better MCMC algorithms in our software could certainly be looked into as future work. But perhaps a more important thing to look at for the MCMC would be a way of better assessing convergence. As we recall, all we have done for this in this thesis is visualize the MCMC results with diagnosis illustrations. This is maybe lacking in rigour even though this is often what is done by MCMC practitioners.

For the SOO ozone data record application of the DLM procedure, the reported standard deviations are an essential component of the DLM model. They completely define the model matrix $\mathbf{V}_i$. However, these standard deviations are known to be fairly unreliable. One option to deal with this is to not include the reported standard deviations in the model at all, and

instead assume that the standard deviation is constant across the data and have this constant number estimated by the MCMC experiment. Mathematically, what we mean is that we set the DLM model matrix V_i equal to some value V for all i, and the V is a quantity that is put in $\boldsymbol{\theta}$.

Another similar option could be to have different values of V for different regions of the time series (i.e. for different i). For example, we could assume V is something different for the SAGE II, OSIRIS, and OMPS portions of the time series as well as in the two overlap periods. This would add 5 variables to $\boldsymbol{\theta}$ however, which could potentially lead to some problems for the convergence of the MCMC chain.

The last option we give, which we think has the most potential, is that we could set $V_i = ak_i$, where we have a as a parameter in $\boldsymbol{\theta}$ and the k_i specified by the modeller. So, this means that the k_i (for all i) should just be a relative quantity of the variance. This would only add one parameter to $\boldsymbol{\theta}$, but the trouble is determining what to specify k_i as. However, estimating k_i is of course theoretically simpler than estimating the variance all together. So, if it is realized that some method to estimate k_i can be devised which is easier than estimating the variances all together for the SOO data record, this could be useful. One simple example could be to choose these weights to be $1/n$, where n is the total number of data points that each MZM is constructed with. This would specify to the DLM, for instance, that the error is much larger for the SAGE II portion of the time series than the rest.

Lastly for potential improvement and future work, a further and more thorough assessment of the DLM procedure could be done. In this thesis, we have developed a statistical procedure and presented the results for an example application. The next step is often to evaluate the procedure in some way to ensure it is suited for a problem. What we did do is note how our results for the example application agree fairly well with the Prais-Winsten procedure, and how theoretically the results are better because they account for more error and do not restrict the shape of the trend terms. Further assessment of the DLM procedure should be included as future work. The literature on methods for evaluating statistical models (in general) could maybe be useful in this endeavour.

References

Alsing, J. (2019). Dlmmc: Dynamical linear model regression for atmospheric time-series analysis. *Journal of Open Source Software*, *4*(37), 1157. Retrieved from https://doi.org/10.21105/joss.01157

Asada, H. (2006). 2.610 identification, estimation, and learning. Retrieved from https://ocw.mit.edu. (License: Creative Commons BY-NC-SA)

Ball, W. T., Alsing, J., Mortlock, D. J., Rozanov, E. V., Tummon, F., & Haigh, J. D. (2017). Reconciling differences in stratospheric ozone composites. *Atmospheric Chemistry and Physics*, *17*(20), 12269–12302. doi:10.5194/acp-17-12269-2017

Ball, W. T., Alsing, J., Mortlock, D. J., Staehelin, J., Haigh, J. D., Peter, T., Tummon, F., Stübi, R., Stenke, A., Anderson, J., Bourassa, A., Davis, S. M., Degenstein, D., Frith, S., Froidevaux, L., Roth, C., Sofieva, V., Wang, R., Wild, J., Yu, P., Ziemke, J. R., & Rozanov, E. V. (2018). Evidence for a continuous decline in lower stratospheric ozone offsetting ozone layer recovery. *Atmospheric Chemistry and Physics*, *18*(2), 1379–1394. doi:10.5194/acp-18-1379-2018

Bourassa, A. E., Roth, C. Z., Zawada, D. J., Rieger, L. A., McLinden, C. A., & Degenstein, D. A. (2018). Drift-corrected odin-osiris ozone product: Algorithm and updated stratospheric ozone trends. *Atmospheric Measurement Techniques*, *11*(1), 489–498. doi:10.5194/amt-11-489-2018

Brown, R., & Hwang, P. (2012). Discrete kalman filter basics. In *Introduction to random signals and applied kalman filtering with matlab exercises* (Chap. 4). CourseSmart Series. Wiley. Retrieved from https://books.google.ca/books?id=z9FiAgAAQBAJ

Damadeo, R. P., Zawodny, J. M., & Thomason, L. W. (2014). Reevaluation of stratospheric ozone trends from sage ii data using a simultaneous temporal and spatial analysis. *Atmospheric Chemistry and Physics*, *14*(24), 13455–13470. doi:10.5194/acp-14-13455-2014

Devore, J., & Berk, K. (2011). *Modern mathematical statistics with applications*. Springer Texts in Statistics. Springer New York. Retrieved from https://books.google.ca/books?id=5PRLUho-YYgC

Dobson, G. M. B. (1968). Forty years' research on atmospheric ozone at oxford: A history. *Appl. Opt.* *7*(3), 387–405. doi:10.1364/AO.7.000387

Farman, J. C., Gardiner, B. G., & Shanklin, J. D. (1985). Large losses of total ozone in antarctica reveal seasonal clox/nox interaction. *Nature*, *315*(6016), 207–210. doi:10.1038/315207a0

Farman, J. C., Gardiner, B. G., & Shanklin, J. D. (2010). Molecular mechanisms of ultraviolet radiation-induced dna damage and repair. *Journal of nucleic acids*, *2010*, 592980. doi:10.4061/2010/592980

Frith, S. M., Kramarova, N. A., Stolarski, R. S., McPeters, R. D., Bhartia, P. K., & Labow, G. J. (2014). Recent changes in total column ozone based on the sbuv version 8.6

merged ozone data set. *Journal of Geophysical Research: Atmospheres*, *119*(16), 9735–9751. doi:10.1002/2014JD021889. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014JD021889

Gebhardt, C., Rozanov, A., Hommel, R., Weber, M., Bovensmann, H., Burrows, J. P., Degenstein, D., Froidevaux, L., & Thompson, A. M. (2014). Stratospheric ozone trends and variability as seen by sciamachy from 2002 to 2012. *Atmospheric Chemistry and Physics*, *14*(2), 831–846. doi:10.5194/acp-14-831-2014

Gilks, W., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain monte carlo in practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis. Retrieved from http://books.google.com/books?id=TRXrMWY%5C_i2IC

Jazwinski, A. (1970). Linear filtering theory. In *Stochastic processes and filtering theory* (Chap. 7, 64, pp. 200–204). Mathematics in science and engineering. New York, NY [u.a.]: Acad. Press. Retrieved from http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+021832242&sourceid=fbw_bibsonomy

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35–45. doi:10.1115/1.3662552. eprint: https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/82/1/35/5518977/35_1.pdf

Kyrölä, E., Laine, M., Sofieva, V., Tamminen, J., Päivärinta, S.-M., Tukiainen, S., Zawodny, J., & Thomason, L. (2013). Combined sage ii-gomos ozone profile data set for 1984-2011 and trend analysis of the vertical distribution of ozone. *Atmospheric Chemistry and Physics*, *13*(21), 10645–10658. doi:10.5194/acp-13-10645-2013

Laeng, A., von Clarmann, T., Stiller, G., Dinelli, B. M., Dudhia, A., Raspollini, P., Glatthor, N., Grabowski, U., Sofieva, V., Froidevaux, L., Walker, K. A., & Zehner, C. (2017). Merged ozone profiles from four mipas processors. *Atmospheric Measurement Techniques*, *10*(4), 1511–1518. doi:10.5194/amt-10-1511-2017

Laine, M., Latva-Pukkila, N., & Kyrölä, E. (2014). Analysing time-varying trends in stratospheric ozone time series using the state space approach. *Atmospheric Chemistry and Physics*, *14*(18), 9707–9725. doi:10.5194/acp-14-9707-2014

Nair, P. J., Godin-Beekmann, S., Kuttippurath, J., Ancellet, G., Goutail, F., Pazmiño, A., Froidevaux, L., Zawodny, J. M., Evans, R. D., Wang, H. J., Anderson, J., & Pastel, M. (2013). Ozone trends derived from the total column and vertical profiles at a northern mid-latitude station. *Atmospheric Chemistry and Physics*, *13*(20), 10373–10384. doi:10.5194/acp-13-10373-2013

Petris, G., Petrone, S., & Campagnoli, P. (2009). Dynamic linear models with R. (Vol. 38, pp. 31–84). doi:10.1007/b135794_2

Petropavlovskikh, I., Godin-Beekmann, S., Hubert, D., Damadeo, R., Hassler, B., & Sofieva, V. (2019). *Sparc/io3c/gaw report on long-term ozone trends and uncertainties in the stratosphere*. 9th assessment report of the SPARC project, published by the International Project Office at DLR-IPA. also: GAW Report No. 241; WCRP Report 17/2018. Retrieved from https://elib.dlr.de/126666/

Prais, S., & Winsten, C. (1954). Trend estimators and serial correlation. *Chicago: Cowles Commission discussion paper*, *383*, 1–12. Retrieved from https://cowles.yale.edu/sites/default/files/files/pub/cdp/s-0383.pdf

Rauch, H. E., Tung, F., & Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, *3*(8), 1445–1450. doi:10.2514/3.3166. eprint: https://doi.org/10.2514/3.3166

Rodgers, C. (2000). *Inverse methods for atmospheric sounding: Theory and practice*. Series on atmospheric, oceanic and planetary physics. World Scientific. Retrieved from https://books.google.ca/books?id=dW-0QgAACAAJ

Savin, N. E., & White, K. J. (1978). Testing for autocorrelation with missing observations. *Econometrica*, *46*(1), 59–67. Retrieved from http://www.jstor.org/stable/1913645

Sofieva, V. F., Kyrölä, E., Laine, M., Tamminen, J., Degenstein, D., Bourassa, A., Roth, C., Zawada, D., Weber, M., Rozanov, A., Rahpoe, N., Stiller, G., Laeng, A., von Clarmann, T., Walker, K. A., Sheese, P., Hubert, D., van Roozendael, M., Zehner, C., Damadeo, R., Zawodny, J., Kramarova, N., & Bhartia, P. K. (2017). Merged sage ii, ozone_cci and omps ozone profile dataset and evaluation of ozone trends in the stratosphere. *Atmospheric Chemistry and Physics*, *17*(20), 12533–12552. doi:10.5194/acp-17-12533-2017

Sorenson, H. (1970). Comparison of kalman, bayesian and maximum likelihood estimation techniques. In C. Leondes (Ed.), *Theory and applications of kalman filtering* (Chap. 6, pp. 138–140). AGARDograph. Retrieved from https://apps.dtic.mil/docs/citations/AD0704306

Steinbrecht, W., Froidevaux, L., Fuller, R., Wang, R., Anderson, J., Roth, C., Bourassa, A., Degenstein, D., Damadeo, R., Zawodny, J., Frith, S., McPeters, R., Bhartia, P., Wild, J., Long, C., Davis, S., Rosenlof, K., Sofieva, V., Walker, K., Rahpoe, N., Rozanov, A., Weber, M., Laeng, A., von Clarmann, T., Stiller, G., Kramarova, N., Godin-Beekmann, S., Leblanc, T., Querel, R., Swart, D., Boyd, I., Hocke, K., Kämpfer, N., Maillard Barras, E., Moreira, L., Nedoluha, G., Vigouroux, C., Blumenstock, T., Schneider, M., Garcia, O., Jones, N., Mahieu, E., Smale, D., Kotkamp, M., Robinson, J., Petropavlovskikh, I., Harris, N., Hassler, B., Hubert, D., & Tummon, F. (2017). An update on ozone profile trends for the period 2000 to 2016. *Atmospheric Chemistry and Physics*, *17*(17), 10675–10690. doi:10.5194/acp-17-10675-2017

West, M., & Harrison, J. (1997). *Bayesian forecasting and dynamic models (2nd ed.)* Berlin, Heidelberg: Springer-Verlag.

WMO. (2018). *Scientific assessment of ozone depletion: 2018 - report and executive summary* (tech. rep. No. Report No. 58). World Meteorological Organization. Retrieved from https://ozone.unep.org/sites/default/files/Assessment_Panel/SAP-2018-Assessment-report.pdf

Wolter, K., & Timlin, M. S. (2011). El niño/southern oscillation behaviour since 1871 as diagnosed in an extended multivariate enso index (mei.ext). *International Journal of Climatology*, *31*(7), 1074–1087. doi:10.1002/joc.2336. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.2336

Appendix A

MLR Least Squares Derivation

Ordinary Least Squares

The sum of squares is given as,

$$S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \tag{A.1}$$
$$= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{A.2}$$

Taking the derivative of this with respect to $\boldsymbol{\beta}$ (using identities in Appendix P) and setting the derivative equal to zero we have,

$$\left(\frac{dS(\boldsymbol{\beta})}{d\boldsymbol{\beta}}\right)^T = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0. \tag{A.3}$$

The solution $\boldsymbol{\beta}^*$ to this equation is

$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{A.4}$$

Generalized Least Squares

Under the transformed model for the GLS technique, we have $\mathbf{GY}$ as $\mathbf{Y}$, and $\mathbf{GX}$ as $\mathbf{X}$ in comparison to the regular model equation where the BLUE is given by OLS estimation. So, we can see that to find the GLS estimate of the regression coefficients, all we need to do is plug in these new matrices into Equation A.4. This gives,

$$\boldsymbol{\beta}^* = ((\mathbf{GX})^T\mathbf{GX})^{-1}(\mathbf{GX})^T\mathbf{Gy} \tag{A.5}$$
$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{G}^T\mathbf{GX})^{-1}\mathbf{X}^T\mathbf{G}^T\mathbf{Gy} \tag{A.6}$$
$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}. \tag{A.7}$$

Alternatively, if we wrote the sum of squares for this GLS case we would have,

$$S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \tag{A.8}$$
$$= (\mathbf{Gy} - \mathbf{GX}\boldsymbol{\beta})^T(\mathbf{Gy} - \mathbf{GX}\boldsymbol{\beta}), \tag{A.9}$$
$$\tag{A.10}$$

and of course minimizing this leads to the same GLS estimate.

Appendix B

Unbiased σ^{2^*} Estimator

The purpose of this appendix is to show that for an MLR model using OLS estimation, the estimate of σ^2 given as,

$$\sigma^{2^*} = \frac{\boldsymbol{\epsilon^*}^{\mathbf{T}}\boldsymbol{\epsilon^*}}{n-(k+1)}, \tag{B.1}$$

is unbiased.

Proof

We have,

$$\begin{aligned}
\sigma^{2^*} &= \frac{\boldsymbol{\epsilon^*}^{\mathbf{T}}\boldsymbol{\epsilon^*}}{n-(k+1)} && \text{(B.2)}\\
&= \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta^*})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta^*})}{n-(k+1)} && \text{(B.3)}\\
&= \frac{(\mathbf{y}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})^T(\mathbf{y}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})}{n-(k+1)}, && \text{(B.4)}
\end{aligned}$$

Plugging in $\mathbf{Y}$ for $\mathbf{y}$ to obtain an estimator of this quantity, which we denote as S^2, gives,

$$\begin{aligned}
S^2 &= \frac{(\mathbf{Y}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y})^T(\mathbf{Y}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y})}{n-(k+1)} && \text{(B.5)}\\
&= \frac{(\mathbf{Y}-\mathbf{H}\mathbf{Y})^T(\mathbf{Y}-\mathbf{H}\mathbf{Y})}{n-(k+1)}, && \text{(B.6)}
\end{aligned}$$

where we define $\mathbf{H}$ as $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. The expectation of this estimator is given as,

$$\mathrm{E}[S^2] = \frac{1}{n-(k+1)}\mathrm{E}[\mathbf{Y}^T\mathbf{Y}-\mathbf{Y}^T\mathbf{H}^T\mathbf{Y}-\mathbf{Y}^T\mathbf{H}\mathbf{Y}+\mathbf{Y}^T\mathbf{H}^T\mathbf{H}\mathbf{Y}]. \tag{B.7}$$

Noting that $\mathbf{H}=\mathbf{H}^T=\mathbf{H}^T\mathbf{H}$, we may write,

$$\mathrm{E}[S^2] = \frac{1}{n-(k+1)}\mathrm{E}[\mathbf{Y}^T\mathbf{Y}-\mathbf{Y}^T\mathbf{H}\mathbf{Y}]. \tag{B.8}$$

The first term is

$$\mathrm{E}[\mathbf{Y}^T\mathbf{Y}] = \mathrm{E}[(\mathbf{X}\boldsymbol{\beta}+\mathbf{e})^T(\mathbf{X}\boldsymbol{\beta}+\mathbf{e})] \tag{B.9}$$
$$= \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + n\sigma^2. \tag{B.10}$$

The second term is

$$\mathrm{E}[\mathbf{Y}^T\mathbf{H}\mathbf{Y}] = \mathrm{E}[(\mathbf{X}\boldsymbol{\beta}+\mathbf{e})^T\mathbf{H}(\mathbf{X}\boldsymbol{\beta}+\mathbf{e})] \tag{B.11}$$
$$= \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{H}\mathbf{X}\boldsymbol{\beta} + \mathrm{tr}(\mathbf{H})\sigma^2 \tag{B.12}$$
$$= \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \mathrm{tr}(\mathbf{H})\sigma^2. \tag{B.13}$$

So, for $\mathrm{E}[S^2]$ we have,

$$\mathrm{E}[S^2] = \frac{\sigma^2}{n-(k+1)}(n - \mathrm{tr}(\mathbf{H})). \tag{B.14}$$

It can be shown that when the number of columns of $\mathbf{X}$ is less than or equal to the number of rows, that the trace of $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is equal to the number of columns of $\mathbf{X}$. We know that for the MLR model the number of columns of $\mathbf{X}$ is $k+1$. Therefore, we have shown that S^2 is an unbiased estimator of σ^2 by showing that

$$\mathrm{E}[S^2] = \sigma^2. \tag{B.15}$$

Note that the condition that the number of columns of $\mathbf{X}$ being less than the number of rows is almost always the case. Lastly, note that although S^2 is an unbiased estimator of σ^2, S is not an unbiased estimator of σ also.

Appendix C

The Minimum Mean Squared Error Statistic

Similarly to the way which the variance of a random variable is defined, as an expectation of a squared deviation, the mean squared error (MSE) of an estimator $\hat{\theta}$ of some parameter of interest θ is defined as $\mathrm{E}[(\hat{\theta} - \theta)^2]$.

Definition 1 *MSE of an estimator $\hat{\theta}$ of some parameter of interest θ:* $\mathrm{E}[(\hat{\theta} - \theta)^2]$

If one is looking for an estimator of θ a reasonable goal might be to seek an estimator where the MSE is smaller than that of every other estimator of θ. This is called the minimum mean square error (MMSE) estimator.

This is oftentimes a too ambitions goal however (Devore and Berk, 2011). Instead, a common restriction that can be made is that the estimator must be "unbiased". This means that $\mathrm{E}[\hat{\theta}] = \theta$, or in words, that the expectation value of the estimator is equal to the parameter it is meant to estimate. So, the MMSE estimator can be found under this restriction. The bias of an estimator is defined as $\mathrm{E}[\hat{\theta}] - \theta$.

Definition 2 *Bias of an estimator:* $\mathrm{E}[\hat{\theta}] - \theta$

Definition 3 *An estimator is unbiased if:* $\mathrm{E}[\hat{\theta}] = \theta$

Notice that if the common equation for the variance of a random variable,

$$Var(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2, \tag{C.1}$$

is used, then the MSE can be expressed in the following form:

$$\mathrm{E}[(\hat{\theta} - \theta)^2] = Var(\hat{\theta} - \theta) + \mathrm{E}[\hat{\theta} - \theta]^2 \tag{C.2}$$

$$= Var(\hat{\theta}) + (\mathrm{E}[\hat{\theta}] - \theta)^2. \tag{C.3}$$

The result is that the MSE is shown to be the sum of the variance of the estimator and the square of the bias of the estimator. So, if the estimator is restricted to be unbiased, the MMSE estimator is the same as the estimator that has minimum variance out of all possible unbiased estimators. This is called the minimum variance unbiased estimator (MVUE). For regression analysis, the Gauss-Markov theorem (Appendix E) considers the MVUE, but only for linear estimators. In this case, the estimate is often called the best linear unbiased estimator (BLUE).

If the parameter of interest is a vector $\boldsymbol{\theta}$ of size n rather than just a single number then the mean squared error is given by,

$$\sum_{i=1}^{n} \mathrm{E}[(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^2]. \tag{C.4}$$

With simple linear algebra, this can be expressed in a few different ways as follows:

$$\sum_{i=1}^{n} \mathrm{E}[(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^2] = \mathrm{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})] \tag{C.5}$$

$$= \mathrm{tr}(\mathrm{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T]) \tag{C.6}$$

$$= \mathrm{E}[\mathrm{tr}((\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T)]. \tag{C.7}$$

The definition of bias for the multivariate case is analogous to the univariate case with it being $\mathrm{E}[\boldsymbol{\theta}] - \boldsymbol{\theta}$.

Note that so far we have treated θ as a scalar quantity and $\boldsymbol{\theta}$ as a vector variable. However, they could also be treated as random quantities and the same definition of mean squared error applies. If this is the case then the definitions of bias can be extended to be $\mathrm{E}[\hat{\theta}] - \mathrm{E}[\theta]$ in the univariate case and $\mathrm{E}[\hat{\boldsymbol{\theta}}] - \mathrm{E}[\boldsymbol{\theta}]$ in the multivariate case.

C.1 MMSE as a Conditional Expectation

In this section, it is proven that if we have observed a random vector $\mathbf{Y}$ as the value $\mathbf{y}$, then the MMSE estimate $\hat{\mathbf{x}}_m$ out of all estimates $\hat{\mathbf{x}}$ of the random vector $\mathbf{X}$ is the same as the conditional expectation,

$$\hat{\mathbf{x}}_m = \mathrm{E}[\mathbf{X}|\mathbf{y}], \tag{C.8}$$

under the restriction that the estimates must be a function of $\mathbf{y}$ (i.e. $\hat{\mathbf{x}} = \mathbf{g}(\mathbf{y})$). Before the proof begins, let us also define $\hat{\mathbf{X}}_m$ as the MMSE estimator and $\hat{\mathbf{X}} = \mathbf{g}(\mathbf{Y})$ as the valid estimators of the random vector $\mathbf{X}$.

Proof

The MSE of the estimate $\hat{\mathbf{x}}$ is given as,

$$MSE = \mathrm{E}[(\hat{\mathbf{x}} - \mathbf{X})^T(\hat{\mathbf{x}} - \mathbf{X})|\mathbf{y}] \tag{C.9}$$

$$= \hat{\mathbf{x}}^T\hat{\mathbf{x}} - \mathrm{E}[\mathbf{X}|\mathbf{y}]^T\hat{\mathbf{x}} - \hat{\mathbf{x}}^T\mathrm{E}[\mathbf{X}|\mathbf{y}] + \mathrm{E}[\mathbf{X}^T\mathbf{X}|\mathbf{y}]. \tag{C.10}$$

Finding the minimum of this to find the MMSE we have,

$$\left(\frac{dMSE}{d\hat{\mathbf{x}}}\right)^T = 2\hat{\mathbf{x}}_m - \mathrm{E}[\mathbf{X}|\mathbf{y}] - \mathrm{E}[\mathbf{X}|\mathbf{y}] = 0. \tag{C.11}$$

Therefore, $\hat{\mathbf{x}}_m$ is given as,

$$\hat{\mathbf{x}}_m = \mathrm{E}[\mathbf{X}|\mathbf{y}]. \tag{C.12}$$

Replacing $\mathbf{Y}$ for $\mathbf{y}$, the corresponding estimator is given as,

$$\hat{\mathbf{X}}_m = \mathrm{E}[\mathbf{X}|\mathbf{Y}]. \tag{C.13}$$

We also note that this estimator is unbiased. We see this because taking the expected value of this random vector gives $\mathrm{E}[\hat{\mathbf{X}}_m] = \mathrm{E}[\mathbf{X}]$ by the law of total expectation. Lastly, the MSE of the estimator $\hat{\mathbf{X}}$ can also be written without any conditioning on $\mathbf{Y}$ as,

$$MSE = \mathrm{E}[(\hat{\mathbf{X}} - \mathbf{X})^T(\hat{\mathbf{X}} - \mathbf{X})] = \mathrm{E}[(\mathbf{g}(\mathbf{Y}) - \mathbf{X})^T(\mathbf{g}(\mathbf{Y}) - \mathbf{X})]. \tag{C.14}$$

Appendix D

Covariance with a Linear Operator

The covariance of a random vector $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{A}$ is some matrix and $\mathbf{X}$ is a random vector, is given as,

$$\mathrm{Cov}[\mathbf{Y}] = \mathbf{A}\mathrm{Cov}[\mathbf{X}]\mathbf{A}^T. \tag{D.1}$$

Proof

$$\begin{aligned}
\mathrm{Cov}[\mathbf{Y}] &= \mathrm{E}[(\mathbf{AX} - \mathbf{A}\mathrm{E}[\mathbf{X}])(\mathbf{AX} - \mathbf{A}\mathrm{E}[\mathbf{X}])^T] && \text{(D.2)}\\
&= \mathrm{E}[\mathbf{A}(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^T\mathbf{A}^T] && \text{(D.3)}\\
&= \mathbf{A}\mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^T]\mathbf{A}^T && \text{(D.4)}\\
&= \mathbf{A}\mathrm{Cov}[\mathbf{X}]\mathbf{A}^T && \text{(D.5)}
\end{aligned}$$

Appendix E

The Gauss-Markov Theorem

The Gauss-Markov theorem is the proof that the OLS estimator of $\boldsymbol{\beta}$ for the MLR model is, out of all possible unbiased estimators that are linear in $\mathbf{Y}$, the unbiased estimator with minimum variance (see Appendix C for more theoretical background on the idea behind creating an unbiased estimator with minimum variance, and why this is good). In this appendix, rather than presenting the proof for the traditional Gauss-Markov theorem, the GLS estimator for a general covariance structure is proven to be the minimum variance unbiased linear estimator instead. The traditional proof can simply be seen from this with $\mathbf{V} = \mathbf{I}$.

Proof

Recall the GLS estimator given as,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}, \tag{E.1}$$

and define another linear in $\mathbf{Y}$ estimator as $\boldsymbol{\beta}' = \mathbf{C}\mathbf{Y}$, where $\mathbf{C} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1} + \mathbf{D}$ and $\mathbf{D}$ is a non-zero matrix. The point of specifying this second estimator like this is to encompasses all possible linear in $\mathbf{Y}$ unbiased estimators.

The expectation value of the $\boldsymbol{\beta}'$ estimator is given as follows:

$$\begin{aligned}
\mathrm{E}[\boldsymbol{\beta}'] &= \mathrm{E}[\mathbf{C}\mathbf{Y}] && \text{(E.2)}\\
&= \mathrm{E}[((\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1} + \mathbf{D})(\mathbf{X}\boldsymbol{\beta} + \mathbf{e})] && \text{(E.3)}\\
&= ((\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1})\mathbf{X}\boldsymbol{\beta} && \text{(E.4)}\\
&= \boldsymbol{\beta} + \mathbf{D}\mathbf{X}\boldsymbol{\beta} && \text{(E.5)}\\
&= (\mathbf{I} + \mathbf{D}\mathbf{X})\boldsymbol{\beta}. && \text{(E.6)}
\end{aligned}$$

Therefore, if the estimator $\boldsymbol{\beta}'$ is to be unbiased, $\mathbf{D}\mathbf{X}$ must be equal to zero.

Now, using the fact that $\mathbf{D}\mathbf{X}$ is equal to zero, the covariance matrix of the $\boldsymbol{\beta}'$ estimator is given as follows:

$$\begin{aligned}
\text{Cov}[\boldsymbol{\beta}'] &= \text{Cov}[\mathbf{CY}] && \text{(E.7)}\\
&= \mathbf{C}\text{Cov}[\mathbf{Y}]\mathbf{C}^T && \text{(E.8)}\\
&= a^2\mathbf{CVC}^T && \text{(E.9)}\\
&= a^2((\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1} + \mathbf{D})\mathbf{V}((\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1} + \mathbf{D})^T && \text{(E.10)}\\
&= a^2((\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{DV})\mathbf{V}^{-1}\mathbf{V}(\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} + \mathbf{D}^T) && \text{(E.11)}\\
&= a^2((\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} + \mathbf{DX}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} + (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{DX})^T + \mathbf{DVD}^T) && \text{(E.12)}\\
&= a^2((\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} + \mathbf{DVD}^T) && \text{(E.13)}\\
&= \text{Cov}[\hat{\boldsymbol{\beta}}] + a^2\mathbf{DVD}^T. && \text{(E.14)}
\end{aligned}$$

Therefore, since $\mathbf{DVD}^T$ is guaranteed to be a positive semidefinite matrix with positive diagonal elements, the variance of all the elements of $\boldsymbol{\beta}'$ exceeds that of $\hat{\boldsymbol{\beta}}$. Therefore, out of all possible linear in $\mathbf{Y}$ unbiased estimators of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$ is the one with minimum variance.

Again, for the OLS estimate where $\mathbf{V} = \mathbf{I}$, the same argument is seen. And lastly, it is important to note that alternatively to the approach here, we could have proven this for the OLS case and it would have then been sufficient to just say that the GLS estimator is also minimum variance because the GLS estimation is just an OLS estimation given the transformation described in Section 2.4.2.1.

Appendix F

Verification of Matrix Inverse for Prais-Winsten Estimation

In this appendix, a verification of the inverse of $\mathbf{V}$, as shown in Equation 2.46 and Equation 2.47, is shown. For a non-singular tridiagonal matrix (a matrix where all elements are zeros except for the three center diagonals) T given as,

$$T = \begin{bmatrix} a_1 & b_1 & & & \\ c_1 & a_2 & b_2 & & \\ & c_2 & \ddots & \ddots & \\ & & \ddots & \ddots & b_{n-1} \\ & & & c_{n-1} & a_n \end{bmatrix}, \tag{F.1}$$

the inverse is given as,

$$(T^{-1})_{ij} = \begin{cases} (-1)^{i+j} b_i b_{i+1} \cdots_{j-1} \theta_{i-1}\phi_{j+1}/\theta_n & i < j \\ \theta_{i-1}\phi_{j+1}/\theta_n & i = j \\ (-1)^{i+j} c_j c_{j+1} \ldots_{i-1} \theta_{j-1}\phi_{i+1}/\theta_n & i > j \end{cases}, \tag{F.2}$$

where the θ_i satisfy,

$$\theta_i = a_i\theta_{i-1} - b_{i-1}c_{i-1}\theta_{i-2} \qquad i = 2, 3, ..., n, \tag{F.3}$$

with initial conditions $\theta_0 = 1$ and $\theta_1 = a_1$, and the ϕ_i satisfy,

$$\phi_i = a_i\phi_{i+1} - b_i c_i \phi_{i+2} \qquad i = n-1, n-2, ..., 2. \tag{F.4}$$

With this, we can work backwards and find the inverse of $\mathbf{V}^{-1}$ to verify that this is equal to the $\mathbf{V}$ given in Equation 2.46. The θ_i and ϕ_i for this particular matrix are all seen to be equal to 1, except for θ_n, which is equal to $1 - \rho^2$. Therefore, it is shown that

$$(\mathbf{V})_{ij} = \begin{cases} (-1)^{i+j}(-\rho)^{j-i}/(1-\rho^2) & i < j \\ 1/(1-\rho^2) & i = j \\ (-1)^{i+j}(-\rho)^{i-j}/(1-\rho^2) & i > j \end{cases} \tag{F.5}$$

$$= \begin{cases} \rho^{j-i}/(1-\rho^2) & i < j \\ 1/(1-\rho^2) & i = j \\ \rho^{i-j}/(1-\rho^2) & i > j \end{cases}, \tag{F.6}$$

which is indeed consistent with the matrix $\mathbf{V}$ given by Equation 2.46.

Appendix G

The Standard Normal Distribution

If a random variable X is a Gaussian distributed random variable with a mean of μ and a variance of σ^2, then the random variable given as,

$$Z = \frac{X - \mu}{\sigma}, \tag{G.1}$$

is the standard normal distribution (which means it is a Gaussian distributed random variable with a mean of 0 and variance of 1).

Proof

Taking the expectation gives,

$$\mathrm{E}[Z] = \frac{\mathrm{E}[X] - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0. \tag{G.2}$$

Taking the variance gives,

$$\begin{aligned} \mathrm{Var}[Z] &= \mathrm{E}[Z^2] - \mathrm{E}[Z]^2 && \text{(G.3)} \\ &= \mathrm{E}[\frac{X^2 - 2\mu X + \mu^2}{\sigma^2}] - 0 && \text{(G.4)} \\ &= \frac{\mathrm{E}[X^2] - 2\mu^2 + \mu^2}{\sigma^2}, && \text{(G.5)} \end{aligned}$$

where $\mathrm{E}[X^2]$ can be found to be,

$$\mathrm{E}[X^2] = \mathrm{Var}[X] + \mathrm{E}[X]^2 = \sigma^2 + \mu^2. \tag{G.6}$$

Therefore we have,

$$\begin{aligned} \mathrm{Var}[Z] &= \frac{\sigma^2 + \mu^2 - 2\mu^2 + \mu^2}{\sigma^2} && \text{(G.7)} \\ &= \frac{\sigma^2}{\sigma^2} && \text{(G.8)} \\ &= 1. && \text{(G.9)} \end{aligned}$$

Appendix H

The t Distribution for MLR Confidence Intervals

A t distributed random variable with v degrees of freedom can be defined by,

$$T = \frac{Z}{\sqrt{\chi_v / v}}, \tag{H.1}$$

where Z is the standard normal random variable and χ_v is a chi-squared random variable with v degrees of freedom. To show that what we have written for the OLS case in Section 2.4.5.1,

$$T = \frac{\hat{\beta}_i - \beta_i^*}{S\sqrt{(\mathbf{X}^T\mathbf{X})_{ii}^{-1}}}, \tag{H.2}$$

is t distributed, we first rearrange so that,

$$T = \frac{\hat{\beta}_i - \beta_i^*}{\sigma\sqrt{(\mathbf{X}^T\mathbf{X})_{ii}^{-1}}} \frac{1}{\sqrt{\frac{S^2}{\sigma^2}}} \tag{H.3}$$

$$= \frac{\hat{\beta}_i - \beta_i^*}{\sigma\sqrt{(\mathbf{X}^T\mathbf{X})_{ii}^{-1}}} \frac{1}{\sqrt{\frac{(n-(k+1))S^2}{(n-(k+1))\sigma^2}}}. \tag{H.4}$$

Now, since $\frac{\hat{\beta}_i - \beta_i^*}{\sigma\sqrt{(\mathbf{X}^T\mathbf{X})_{ii}^{-1}}}$ is a standard normal random variable Z, what remains is to show that $\frac{(n-(k+1))S^2}{\sigma^2}$ is a chi-squared random variable with $n-(k+1)$ degrees of freedom. This is the case, and will not be shown in this thesis as it is fairly involved.

For the GLS case now, we implied in Section 2.4.5.1 that the associated t distributed random variable is given as,

$$T = \frac{\hat{\beta}_i - \beta_i^*}{S_e\sqrt{(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})_{ii}^{-1}}}, \tag{H.5}$$

where S_e is the estimator of σ_e. So likewise, we can show that

$$T = \frac{\hat{\beta}_i - \beta_i^*}{\sigma_e \sqrt{(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}_{ii}}} \frac{1}{\sqrt{\frac{S_e^2}{\sigma^2}}} \tag{H.6}$$

$$= \frac{\hat{\beta}_i - \beta_i^*}{\sigma_e \sqrt{(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}_{ii}}} \frac{1}{\sqrt{\frac{(n-2)S_e^2}{(n-2)\sigma_e^2}}}, \tag{H.7}$$

and then it remains to show that $\frac{(n-2)S_e^2}{\sigma_e^2}$ is a chi-squared random variable with $n-2$ degrees of freedom.

Appendix I

Matrix Inversion Identities

Woodbury Matrix Identity

$$(\mathbf{A}+\mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1}+\mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1} \tag{I.1}$$

Matrix Addition Inversion Identity

$$(\mathbf{A}+\mathbf{B})^{-1} = \mathbf{A}^{-1} - (\mathbf{A}+\mathbf{B})^{-1}\mathbf{BA}^{-1} \tag{I.2}$$

Appendix J

RLS Algorithm Alternate Form

In this appendix, the suitable matrix $\mathbf{K_i}$ for the RLS algorithm in the following form:

$$\boldsymbol{\beta}_{\mathbf{i}}^{*} = \boldsymbol{\beta}_{\mathbf{i-1}}^{*} + \mathbf{K_i}(y_i - \boldsymbol{\varphi_i}^T \boldsymbol{\beta}_{\mathbf{i-1}}^{*}), \tag{J.1}$$

is found to be

$$\mathbf{K_i} = \frac{\mathbf{P_{i-1}}\boldsymbol{\varphi_i}}{1 + \boldsymbol{\varphi_i}^T \mathbf{P_{i-1}} \boldsymbol{\varphi_i}}. \tag{J.2}$$

Proof

Using the first RLS algorithm presented in this thesis of Equations 4.11, 4.8, and that $\boldsymbol{\beta}_i^{*} = \mathbf{P}_i \mathbf{B}_i$, we find that

$$\boldsymbol{\beta}_{\mathbf{i}}^{*} - \boldsymbol{\beta}_{\mathbf{i-1}}^{*} = (\mathbf{P_{i-1}} - \frac{\mathbf{P_{i-1}}\boldsymbol{\varphi_i}\boldsymbol{\varphi_i}^T\mathbf{P_{i-1}}}{1 + \boldsymbol{\varphi_i}^T \mathbf{P_{i-1}} \boldsymbol{\varphi_i}})(\mathbf{B_{i-1}} + y_i\boldsymbol{\varphi_i}) - \mathbf{P_{i-1}}\mathbf{B_{i-1}} \tag{J.3}$$

$$= \mathbf{P_{i-1}} y_i \boldsymbol{\varphi_i} - \frac{\mathbf{P_{i-1}}\boldsymbol{\varphi_i}\boldsymbol{\varphi_i}^T\mathbf{P_{i-1}}}{1 + \boldsymbol{\varphi_i}^T \mathbf{P_{i-1}} \boldsymbol{\varphi_i}}(\mathbf{B_{i-1}} + y_i\boldsymbol{\varphi_i}) \tag{J.4}$$

$$= \frac{\mathbf{P_{i-1}} y_i \boldsymbol{\varphi_i}(1 + \boldsymbol{\varphi_i}^T \mathbf{P_{i-1}} \boldsymbol{\varphi_i}) - \mathbf{P_{i-1}}\boldsymbol{\varphi_i}\boldsymbol{\varphi_i}^T\mathbf{P_{i-1}}(\mathbf{B_{i-1}} + y_i\boldsymbol{\varphi_i})}{1 + \boldsymbol{\varphi_i}^T \mathbf{P_{i-1}} \boldsymbol{\varphi_i}} \tag{J.5}$$

$$= \frac{\mathbf{P_{i-1}} y_i \boldsymbol{\varphi_i} - \mathbf{P_{i-1}}\boldsymbol{\varphi_i}\boldsymbol{\varphi_i}^T\mathbf{P_{i-1}}\mathbf{B_{i-1}}}{1 + \boldsymbol{\varphi_i}^T \mathbf{P_{i-1}} \boldsymbol{\varphi_i}} \tag{J.6}$$

$$= \frac{\mathbf{P_{i-1}}\boldsymbol{\varphi_i}}{1 + \boldsymbol{\varphi_i}^T \mathbf{P_{i-1}} \boldsymbol{\varphi_i}}(y_i - \boldsymbol{\varphi_i}^T \boldsymbol{\beta}_{\mathbf{i-1}}^{*}). \tag{J.7}$$

Equation J.7 is in the form of Equation J.1, therefore we see that we have,

$$\mathbf{K_i} = \frac{\mathbf{P_{i-1}}\boldsymbol{\varphi_i}}{1 + \boldsymbol{\varphi_i}^T \mathbf{P_{i-1}} \boldsymbol{\varphi_i}}. \tag{J.8}$$

Appendix K

RLS Cost Function

In this appendix, we prove the RLS theorem that the minima of the cost function given as,

$$F_n(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\varphi}_i^T\boldsymbol{\beta})^2 + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0^*)^T\mathbf{P}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0^*) \tag{K.1}$$

is at $\boldsymbol{\beta} = \boldsymbol{\beta}_n^*$.

Proof

In this proof, we start by showing that the theorem is true for $n = 1$. We then use this result to manipulate the cost function so that we can get $F_n(\boldsymbol{\beta})$ in a desired form. Then, we show that the minimum of $F_n(\boldsymbol{\beta})$ is at $\boldsymbol{\beta}_n^*$, proving the theorem for any n.

First we write $F_1(\boldsymbol{\beta})$ as,

$$F_1(\boldsymbol{\beta}) = \frac{1}{2}(y_1 - \boldsymbol{\varphi}_1\boldsymbol{\beta})^2 + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0^*)^T\mathbf{P}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0^*). \tag{K.2}$$

Differentiating this with the matrix derivative identities in Appendix P and setting the derivative equal to zero we have,

$$\left(\frac{\partial F_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T = -\boldsymbol{\varphi}_1(y_1 - \boldsymbol{\varphi}_1^T\boldsymbol{\beta}') + \mathbf{P}_0^{-1}(\boldsymbol{\beta}' - \boldsymbol{\beta}_0^*) = 0, \tag{K.3}$$

where we define $\boldsymbol{\beta}'$ as the minimum location of $F_1(\boldsymbol{\beta})$. Therefore, we have,

$$\boldsymbol{\beta}' = (\boldsymbol{\varphi}_1\boldsymbol{\varphi}_1^T + \mathbf{P}_0^{-1})^{-1}(\mathbf{P}_0^{-1}\boldsymbol{\beta}_0^* + \boldsymbol{\varphi}_1 y_1). \tag{K.4}$$

There is a little bit of work left in order to show that this is equivalent to $\boldsymbol{\beta}_1^*$. We will show that this is equivalent to $\boldsymbol{\beta}_1^* = \mathbf{P}_1\mathbf{B}_1$, which is the first version of the RLS algorithm presented in Section 4.1. Recall from Equation 4.11 that $\mathbf{P}_1 = (\boldsymbol{\varphi}_1\boldsymbol{\varphi}_1^T + \mathbf{P}_0^{-1})^{-1}$. So, Equation K.4 can be written as,

$$\boldsymbol{\beta}' = \mathbf{P}_1(\mathbf{P}_0^{-1}\boldsymbol{\beta}_0^* + \boldsymbol{\varphi}_1 y_1). \tag{K.5}$$

Also, recall from Equation 4.8 that $\mathbf{B}_1 = \mathbf{B}_0 + \boldsymbol{\varphi}_1 y_1$ and that we have $\boldsymbol{\beta}_0^* = \mathbf{P}_0\mathbf{B}_0$ (or $\mathbf{B}_0 = \mathbf{P}_0^{-1}\boldsymbol{\beta}_0^*$). Putting these together we have,

$$\mathbf{B}_1 = \mathbf{P}_0^{-1}\boldsymbol{\beta}_0^* + \boldsymbol{\varphi}_1 y_1. \tag{K.6}$$

Therefore, we see that Equation K.5 is

$$\boldsymbol{\beta}' = \mathbf{P}_1\mathbf{B}_1 = \boldsymbol{\beta}_1^*, \tag{K.7}$$

and so we have proven the theorem for the $n = 1$ case.

To extend this proof to any n, we start by Taylor series expanding $F_1(\boldsymbol{\beta})$ about $\boldsymbol{\beta}_1^*$ (see Appendix S for details on the multivariate Taylor series expansion). This gives,

$$\begin{aligned} F_1(\boldsymbol{\beta}) = \frac{1}{2}(y_1 - \boldsymbol{\varphi}_1\boldsymbol{\beta}_1^*)^2 + \frac{1}{2}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^*)^T\mathbf{P}_0^{-1}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^*) + \frac{\partial F_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_1^*}(\boldsymbol{\beta} - \boldsymbol{\beta}_1^*) \\ + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_1^*)^T\frac{\partial}{\partial \boldsymbol{\beta}}\left(\frac{\partial F_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_1^*}(\boldsymbol{\beta} - \boldsymbol{\beta}_1^*). \end{aligned} \tag{K.8}$$

We have $\frac{\partial F_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_1^*}$ equal to zero because $\boldsymbol{\beta}_1^*$ is the minima location of $F_1(\boldsymbol{\beta})$, as we have just shown. To calculate the second order derivative, we have,

$$\left(\frac{\partial F_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T = -\boldsymbol{\varphi}_1(y_1 - \boldsymbol{\varphi}_1^T\boldsymbol{\beta}) + \mathbf{P}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0^*), \tag{K.9}$$

then we have,

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}}\left(\frac{\partial F_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T &= \boldsymbol{\varphi}_1\boldsymbol{\varphi}_1^T + \mathbf{P}_0^{-1} && \text{(K.10)} \\ &= \mathbf{P}_1^{-1}. && \text{(K.11)} \end{aligned}$$

So, if we define the constant term in the Taylor series expansion version of $F_1(\boldsymbol{\beta})$ as

$$C_1 = \frac{1}{2}(y_1 - \boldsymbol{\varphi}_1\boldsymbol{\beta}_1^*)^2 + \frac{1}{2}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^*)^T\mathbf{P}_0^{-1}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^*), \tag{K.12}$$

then Equation K.8 becomes

$$F_1(\boldsymbol{\beta}) = C_1 + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_1^*)^T\mathbf{P}_1^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_1^*). \tag{K.13}$$

This is the exact function of $F_1(\boldsymbol{\beta})$ and not an approximation because $F_1(\boldsymbol{\beta})$ is clearly quadratic and higher order derivatives are zero. To write $F_2(\boldsymbol{\beta})$ now, we note that our cost function in Equation K.1 has the following property:

$$F_n(\boldsymbol{\beta}) = F_{n-1}(\boldsymbol{\beta}) + \frac{1}{2}(y_n - \boldsymbol{\varphi}_n^T\boldsymbol{\beta})^2. \tag{K.14}$$

So we write $F_2(\boldsymbol{\beta})$ as

$$F_2(\boldsymbol{\beta}) = C_1 + \frac{1}{2}(y_2 - \boldsymbol{\varphi}_2^T\boldsymbol{\beta})^2 + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_1^*)^T\mathbf{P}_1^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_1^*). \tag{K.15}$$

We note that this equation closely resembles Equation K.2 for $F_1(\boldsymbol{\beta})$. So, we can see then that if we were to minimize this function we would find the minimum to similarly be at $\boldsymbol{\beta}_2^*$.

Furthermore, we see that if we perform the Taylor Expansion of $F_2(\boldsymbol{\beta})$ about $\boldsymbol{\beta}_2^*$ and then add the $\frac{1}{2}(y_3 - \boldsymbol{\varphi}_3^T\boldsymbol{\beta})^2$ term exactly as before then we get the following for $F_3(\boldsymbol{\beta})$:

$$F_3(\boldsymbol{\beta}) = C_1 + C_2 + \frac{1}{2}(y_3 - \boldsymbol{\varphi}_3^T\boldsymbol{\beta})^2 + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_2^*)^T\mathbf{P}_2^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_2^*). \tag{K.16}$$

This can be done indefinitely, so at any index n we get,

$$F_n(\boldsymbol{\beta}) = \sum_{i-1}^{n-1} C_i + \frac{1}{2}(y_n - \boldsymbol{\varphi}_n^T\boldsymbol{\beta})^2 + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{n-1}^*)^T\mathbf{P}_{n-1}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_{n-1}^*). \tag{K.17}$$

The constants do not matter since this is a cost function, but it is easy to see that they are given as,

$$C_i = \frac{1}{2}(y_i - \boldsymbol{\varphi}_i^T\boldsymbol{\beta}_i^*)^2 + \frac{1}{2}(\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_{i-1}^*)^T\mathbf{P}_{i-1}^{-1}(\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_{i-1}^*). \tag{K.18}$$

At this point, the proof is pretty much done. Because we know from the math at the beginning of this proof for minimizing $F_1(\boldsymbol{\beta})$, that the minimum location of this function would be found to similarly be at $\boldsymbol{\beta}_n^*$, proving our theorem. We show this once more anyway. Starting with the derivative as,

$$\frac{\partial F_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\boldsymbol{\varphi}_n(y_n - \boldsymbol{\varphi}_n^T\boldsymbol{\beta}') + \mathbf{P}_{n-1}^{-1}(\boldsymbol{\beta}' - \boldsymbol{\beta}_{n-1}^*) = 0. \tag{K.19}$$

Therefore, we have,

$$\boldsymbol{\beta}' = (\boldsymbol{\varphi}_n\boldsymbol{\varphi}_n^T + \mathbf{P}_{n-1}^{-1})^{-1}(\mathbf{P}_{n-1}^{-1}\boldsymbol{\beta}_{n-1}^* + \boldsymbol{\varphi}_n y_n). \tag{K.20}$$

Recall from Equation 4.11 that $\mathbf{P}_n = (\boldsymbol{\varphi}_n\boldsymbol{\varphi}_n^T + \mathbf{P}_{n-1}^{-1})^{-1}$. So, we can write Equation K.20 as,

$$\boldsymbol{\beta}' = \mathbf{P}_n(\mathbf{P}_{n-1}^{-1}\boldsymbol{\beta}_{n-1}^* + \boldsymbol{\varphi}_n y_n). \tag{K.21}$$

Also, recall from Equation 4.8 that $\mathbf{B}_n = \mathbf{B}_{n-1} + \boldsymbol{\varphi}_n y_n$ and that we have $\boldsymbol{\beta}_{n-1}^* = \mathbf{P}_{n-1}\mathbf{B}_{n-1}$ ($\mathbf{B}_{n-1} = \mathbf{P}_{n-1}^{-1}\boldsymbol{\beta}_{n-1}^*$). Putting these together gives

$$\mathbf{B}_n = \mathbf{P}_{n-1}^{-1}\boldsymbol{\beta}_{n-1}^* + \boldsymbol{\varphi}_n y_n. \tag{K.22}$$

Therefore we see that Equation K.21 is the same as

$$\boldsymbol{\beta}' = \mathbf{P}_n\mathbf{B}_n = \boldsymbol{\beta}_n^*, \tag{K.23}$$

and so it has been proven that the minimum of the function $F_n(\boldsymbol{\beta})$ is indeed $\boldsymbol{\beta}_n^*$.

Appendix L

Weighted RLS

In this appendix, we show the modification to the RLS algorithm considering a covariance structure in the errors of the MLR model. So, where the RLS algorithm we show in Section 4.1 is related to an OLS estimate, this weighted RLS algorithm in this appendix is related to a GLS estimate.

By the nature of the RLS algorithm, we can not assume a completely general matrix error covariance structure as we have for the GLS estimate. Instead, we assume a diagonal covariance structure of the errors such that

$$\mathbf{V} = \begin{bmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & V_n \end{bmatrix}. \tag{L.1}$$

Note that rather than factoring out a constant a^2 like what was done in this thesis with the GLS theory, we keep this constant inside $\mathbf{V}$.

Now, the exact same procedure that was carried out in Section 4.1.1 in deriving the RLS algorithm can be followed. This time we just have the GLS estimation formula,

$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}, \tag{L.2}$$

rather than the OLS estimation formula, and so we define $\mathbf{P}$ and $\mathbf{B}$ as $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ and $\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ respectively instead. This exercise is left to the reader. The resulting recurrence relations for the weighted RLS algorithm are given as,

$$\boldsymbol{\beta}_{\mathbf{i}}^* = \boldsymbol{\beta}_{\mathbf{i-1}}^* + \frac{\mathbf{P_{i-1}}\boldsymbol{\varphi}_{\mathbf{i}}}{V_i + \boldsymbol{\varphi}_{\mathbf{i}}{}^T\mathbf{P_{i-1}}\boldsymbol{\varphi}_{\mathbf{i}}}(y_i - \boldsymbol{\varphi}_{\mathbf{i}}{}^T\boldsymbol{\beta}_{\mathbf{i-1}}^*) \tag{L.3}$$

$$\mathbf{P_i} = \mathbf{P_{i-1}} - \frac{\mathbf{P_{i-1}}\boldsymbol{\varphi}_{\mathbf{i}}\boldsymbol{\varphi}_{\mathbf{i}}{}^T\mathbf{P_{i-1}}}{V_i + \boldsymbol{\varphi}_{\mathbf{i}}{}^T\mathbf{P_{i-1}}\boldsymbol{\varphi}_{\mathbf{i}}}, \tag{L.4}$$

where we see that the only thing that changes from the RLS algorithm presented in Section 4.1 is the V_i in place of what was a 1 before.

Lastly, for this weighted RLS algorithm, we can see that the modification to the RLS cost function presented in Section 4.1 is given as,

$$F_n(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{V_i}(y_i - \boldsymbol{\varphi}_i^T\boldsymbol{\beta})^2 + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0^*)^T\mathbf{P}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0^*). \tag{L.5}$$

Appendix M

DLM Model Equations

It is often mentioned in this thesis that for the general DLM, the random vector $\mathbf{X}_i$ that represents the state vector can be written in terms of only the random vectors $\mathbf{X}_0$ and $\mathbf{w}_s$'s where $s \leq i$. This is obvious just by looking at the evolution equation. In this appendix, this, along with a similar result for $\mathbf{Y}_i$, is developed more firmly.

If we continually apply the evolution equation for the general DLM (Equation 4.29) starting with $\mathbf{X}_0$ we have,

$$\mathbf{X}_1 = \mathbf{G}_1\mathbf{X}_0 + \mathbf{w}_1 \tag{M.1}$$
$$\mathbf{X}_2 = \mathbf{G}_2(\mathbf{G}_1\mathbf{X}_0 + \mathbf{w}_1) + \mathbf{w}_2 \tag{M.2}$$
$$\mathbf{X}_3 = \mathbf{G}_3(\mathbf{G}_2(\mathbf{G}_1\mathbf{X}_0 + \mathbf{w}_1) + \mathbf{w}_2) + \mathbf{w}_3. \tag{M.3}$$

It can then be seen that in general this leads to,

$$\mathbf{X}_i = (\prod_{j=1}^{i}\mathbf{G}_j)\mathbf{X}_0 + (\prod_{j=2}^{i}\mathbf{G}_j)\mathbf{w}_1 + (\prod_{j=3}^{i}\mathbf{G}_j)\mathbf{w}_2 + ... + \mathbf{G}_i\mathbf{w}_{i-1} + \mathbf{w}_i \tag{M.4}$$
$$= (\prod_{j=1}^{i}\mathbf{G}_j)\mathbf{X}_0 + \sum_{k=2}^{i+1}((\prod_{j=k}^{i}\mathbf{G}_j)\mathbf{w}_{k-1}). \tag{M.5}$$

Also, from the general DLM observation equation (Equation 4.28) it can be seen that,

$$\mathbf{Y}_i = F_i\mathbf{X}_i + \mathbf{v}_i \tag{M.6}$$
$$= F_i((\prod_{j=1}^{i}\mathbf{G}_j)\mathbf{X}_0 + \sum_{k=2}^{i+1}((\prod_{j=k}^{i}\mathbf{G}_j)\mathbf{w}_{k-1})) + \mathbf{v}_i. \tag{M.7}$$

So, we also have the random vector $\mathbf{Y}_i$ written in terms of only $\mathbf{X}_0$, $\mathbf{w}_s$'s where $s \leq i$, and $\mathbf{v}_i$.

It is important to note that if our initial condition has $\mathrm{E}[\mathbf{X}_0] = 0$ (which is always done in practice) then $\mathrm{E}[\mathbf{Y}_i] = 0$ and $\mathrm{E}[\mathbf{X}_i] = 0$. Also, since we have defined the $\mathbf{w}$'s and $\mathbf{v}$'s to be Gaussian, if $\mathbf{X}_0$ is also Gaussian (which is what we always chose in practice), then the $\mathbf{Y}$'s and $\mathbf{X}$'s are also Gaussian since they are a linear combination of Gaussian distributed random vectors.

As a final note, similarly to the above equations, if we are given the random vector $\mathbf{X}_s$ where $s < i$ then we can write $\mathbf{X}_i$ and $\mathbf{Y}_i$ as

$$\mathbf{X}_i = (\prod_{j=s+1}^{i} \mathbf{G}_j)\mathbf{X}_s + \sum_{k=s+2}^{i+1} ((\prod_{j=k}^{i} \mathbf{G}_j)\mathbf{w}_{k-1}) \tag{M.8}$$

$$\mathbf{Y}_i = F_i((\prod_{j=s+1}^{i} \mathbf{G}_j)\mathbf{X}_s + \sum_{k=s+2}^{i+1} ((\prod_{j=k}^{i} \mathbf{G}_j)\mathbf{w}_{k-1})) + \mathbf{v}_i. \tag{M.9}$$

Appendix N

Bayes Theorem with Gaussian Statistics

In this appendix, we prove that for a Gaussian distributed prior distribution and a Gaussian likelihood function, the posterior distribution resulting from Bayes theorem is Gaussian. This is done for both the univariate and multivariate Gaussian cases. These are well-known results and are typically described in terms of the term "conjugate priors". Furthermore, once this is proved the values of the posterior mean and variance/covariance are found.

We do this for likelihood functions stemming from the following equations. For the univariate case:

$$Y = cX + v, \qquad v \sim N[0, B], \tag{N.1}$$

where c is some constant, and for the multivariate case:

$$\mathbf{Y} = \mathbf{K}\mathbf{X} + \mathbf{v}, \qquad v \sim N[0, \mathbf{B}], \tag{N.2}$$

where $\mathbf{K}$ is some matrix. Note that this are more typically seen for $c = 1$ and $\mathbf{K} = \mathbf{I}$. The linear class of functions used here (i.e. cX apposed to a general $g(X)$) is the only set of functions where the results shown in this appendix hold.

Univariate

The posterior probability density $p(x|y)$ is given with Bayes theorem as,

$$p(x|y) \propto \mathcal{L}(x|y)p(x), \tag{N.3}$$

where $p(x)$ is Gaussian distributed with mean d and variance D and the likelihood function $\mathcal{L}(x|y)$ is written as a Gaussian distribution with mean of cx and variance B. We have,

$$\ln(p(x|y)) = \ln(L(x|y)) + \ln(p(x)) + \ln(c_1) \tag{N.4}$$

$$= -\frac{(y - cx)^2}{2B} - \frac{(x - d)^2}{2D} - c_2, \tag{N.5}$$

where c_1 and c_2 are constants. Since this is quadratic in x it must be possible to write

$$\ln(p(x|y)) = -\frac{(x - \mu)^2}{2V} + c_3, \tag{N.6}$$

where c_3 is another constant. This is true because the parameters μ, V, and c_3 can be chosen to make it so that this is equal to Equation N.5. More specifically, first V can be chosen so that the x^2 terms in equations N.5 and N.6 are equal, then μ can be chosen so that the linear in x terms are equal, then the density function can be normalized for the constant term. Since $p(x|y)$ can be written as it is in Equation N.6 it must be a Gaussian distribution. This

gives us the first result of this appendix that the posterior distribution for the univariate case is Gaussian given a Gaussian likelihood and prior.

Now, the mean μ and variance V of the posterior distribution will be found as described above. By equating the x^2 terms,

$$-\frac{x^2}{2B}-\frac{x^2}{2D}=-\frac{x^2}{2V}, \tag{N.7}$$

we find that the variance V of the posterior density is given as,

$$V=\frac{DB}{D+B}. \tag{N.8}$$

By equating the linear terms in x,

$$-\frac{2cyx}{2B}-\frac{2dx}{2D}=-\frac{2\mu x}{2V}, \tag{N.9}$$

we find the mean μ of the posterior density is given as,

$$\mu=\left(\frac{cy}{B}+\frac{d}{D}\right)V \tag{N.10}$$

$$=\left(\frac{cy}{B}+\frac{d}{D}\right)\frac{DB}{D+B}. \tag{N.11}$$

This mean μ can also be expressed in another form by noting that

$$\mu=d-d+\frac{Vcy}{B}+\frac{Vd}{D} \tag{N.12}$$

$$=d+d\left(\frac{V}{D}-1\right)+\frac{Vcy}{B} \tag{N.13}$$

$$=d+d\left(\frac{B}{D+B}-\frac{D+B}{D+B}\right)+\frac{Dcy}{D+B} \tag{N.14}$$

$$=d+d\left(\frac{-D}{D+B}\right)+\frac{Dcy}{D+B} \tag{N.15}$$

$$=d+\frac{D}{D+B}(cy-d). \tag{N.16}$$

The from of Equation N.16 is arguably more intuitive since we can think of it as the mean of the prior d plus some additional term that takes into account y.

In this univariate section of this appendix, we have proved that a Gaussian likelihood and Gaussian prior result in a Gaussian posterior and we have given equations for the resulting mean and variance of the posterior.

Multivariate

This section follows the steps of the previous univariate section very closely. We have the posterior probability density $p(\mathbf{x}|\mathbf{y})$ given by Bayes theorem as,

$$p(\mathbf{x}|\mathbf{y}) \propto L(\mathbf{x}|\mathbf{y})p(\mathbf{x}), \tag{N.17}$$

where $p(\mathbf{x})$ is Gaussian distributed with mean $\mathbf{d}$ and covariance matrix $\mathbf{D}$ and the likelihood function $L(\mathbf{x}|\mathbf{y})$ is written as a Gaussian distribution with mean $\mathbf{Kx}$ and covariance matrix $\mathbf{B}$. We have,

$$\ln(p(\mathbf{x}|\mathbf{y})) = \ln(L(\mathbf{x}|\mathbf{y})) + \ln(p(\mathbf{x})) + \ln(c_1) \tag{N.18}$$

$$= -\frac{1}{2}(\mathbf{y}-\mathbf{Kx})^T\mathbf{B}^{-1}(\mathbf{y}-\mathbf{Kx}) - -\frac{1}{2}(\mathbf{x}-\mathbf{d})^T\mathbf{D}^{-1}(\mathbf{x}-\mathbf{d}) - c_2, \tag{N.19}$$

where c_1 and c_2 are constants. Since this is quadratic in $\mathbf{x}$ it must be possible to write

$$\ln(p(\mathbf{x}|\mathbf{Y}=\mathbf{y})) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + c_3, \tag{N.20}$$

where c_3 is another constant. This is true because the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and c_3 can be chosen to make it so that this is equal to Equation N.19. More specifically, first $\boldsymbol{\Sigma}$ can be chosen so that the $\mathbf{x}^T\mathbf{x}$ terms in equations N.19 and N.20 are equal, then $\boldsymbol{\mu}$ can be chosen so that the linear terms in $\mathbf{x}$ are equal, then the density function can be normalized for the constant term. Since $p(\mathbf{x}|\mathbf{y})$ can be written as it is in Equation N.20 it must be a multivariate Gaussian distribution, giving us the result that Gaussian distribution Likelihood and priors result in a Gaussian distributed posterior in the Multivariate case as well.

Now, the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ will be found as described above. By equating the $\mathbf{x}^T\mathbf{x}$ terms,

$$-\frac{1}{2}\mathbf{x}^T\mathbf{K}^T\mathbf{B}^{-1}\mathbf{Kx} - \frac{1}{2}\mathbf{x}^T\mathbf{D}^{-1}\mathbf{x} = -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}, \tag{N.21}$$

we find that the covariance matrix $\boldsymbol{\Sigma}$ of the posterior density is

$$\boldsymbol{\Sigma} = (\mathbf{D}^{-1} + \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K})^{-1}. \tag{N.22}$$

By equating linear in $\mathbf{x}$ terms (equating terms linear in $\mathbf{x}^T$ yields the same result),

$$-\frac{1}{2}\mathbf{y}^T\mathbf{B}^{-1}\mathbf{Kx} - \frac{1}{2}\mathbf{d}^T\mathbf{D}^{-1}\mathbf{x} = -\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}, \tag{N.23}$$

and taking the transpose,

$$\mathbf{K}^T\mathbf{B}^{-1}\mathbf{y} + \mathbf{D}^{-1}\mathbf{d} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \tag{N.24}$$

we find that the mean $\boldsymbol{\mu}$ of the posterior density is

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}(\mathbf{K}^T\mathbf{B}^{-1}\mathbf{y} + \mathbf{D}^{-1}\mathbf{d}) \tag{N.25}$$

$$= (\mathbf{D}^{-1} + \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K})^{-1}(\mathbf{K}^T\mathbf{B}^{-1}\mathbf{y} + \mathbf{D}^{-1}\mathbf{d}). \tag{N.26}$$

Now, different forms of the mean and covariance given in Equations N.26 and N.22 can be expressed. Using the Woodbury matrix identity in Appendix I, the covariance can be written alternatively as,

$$\begin{aligned}\boldsymbol{\Sigma} &= (\mathbf{D}^{-1} + \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K})^{-1} && \text{(N.27)}\\ &= \mathbf{D} - \mathbf{D}\mathbf{K}^T(\mathbf{B} + \mathbf{K}\mathbf{D}\mathbf{K}^T)^{-1}\mathbf{K}\mathbf{D}. && \text{(N.28)}\end{aligned}$$

The mean can be written alternatively as,

$$\begin{aligned}\boldsymbol{\mu} &= (\mathbf{D}^{-1} + \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K})^{-1}(\mathbf{K}^T\mathbf{B}^{-1}\mathbf{y} + \mathbf{D}^{-1}\mathbf{d}) && \text{(N.29)}\\ &= \mathbf{d} - \mathbf{d} + (\mathbf{D}^{-1} + \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K})^{-1}(\mathbf{K}^T\mathbf{B}^{-1}\mathbf{y} + \mathbf{D}^{-1}\mathbf{d}) && \text{(N.30)}\\ &= \mathbf{d} + (\mathbf{D}^{-1} + \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K})^{-1}(\mathbf{K}^T\mathbf{B}^{-1}\mathbf{y} + \mathbf{D}^{-1}\mathbf{d} - (\mathbf{D}^{-1} + \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K})\mathbf{d}) && \text{(N.31)}\\ &= \mathbf{d} + (\mathbf{D}^{-1} + \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K})^{-1}(\mathbf{K}^T\mathbf{B}^{-1}\mathbf{y} - \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K}\mathbf{d}) && \text{(N.32)}\\ &= \mathbf{d} + (\mathbf{D}^{-1} + \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K})^{-1}\mathbf{K}^T\mathbf{B}^{-1}(\mathbf{y} - \mathbf{K}\mathbf{d}). && \text{(N.33)}\end{aligned}$$

Now, Equation N.33 can be converted to yet another form which is typically more computationally efficient to calculate. If we have any matrices $\mathbf{W}, \mathbf{Z}, \mathbf{X}, \mathbf{Y}$ and $\mathbf{W}\mathbf{Z} = \mathbf{X}\mathbf{Y}$ where $\mathbf{X}$ and $\mathbf{Z}$ are invertible so that we also have $\mathbf{X}^{-1}\mathbf{W} = \mathbf{Y}\mathbf{Z}^{-1}$. Then, if we define $\mathbf{X}^{-1}$ and $\mathbf{W}$ for the problem at hand to be

$$\mathbf{X}^{-1} = (\mathbf{D}^{-1} + \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K})^{-1}, \tag{N.34}$$

and

$$\mathbf{W} = \mathbf{K}^T\mathbf{B}^{-1}, \tag{N.35}$$

and then arbitrarily define $\mathbf{Y}$ to be $\mathbf{D}\mathbf{K}^T$, we see that we must have $\mathbf{Z} = \mathbf{B} + \mathbf{K}\mathbf{D}\mathbf{K}^T$ if $\mathbf{W}\mathbf{Z} = \mathbf{X}\mathbf{Y}$ is to hold. So, we may write $\mathbf{X}^{-1}\mathbf{W} = \mathbf{Y}\mathbf{Z}^{-1}$ for the problem at hand as,

$$(\mathbf{D}^{-1} + \mathbf{K}^T\mathbf{B}^{-1}\mathbf{K})^{-1}\mathbf{K}^T\mathbf{B}^{-1} = \mathbf{D}\mathbf{K}^T(\mathbf{B} + \mathbf{K}\mathbf{D}\mathbf{K}^T)^{-1}. \tag{N.36}$$

So, from this we see that the another form of the mean $\boldsymbol{\mu}$, which is most typically used since is it the most computationally efficient, is given as,

$$\boldsymbol{\mu} = \mathbf{d} + \mathbf{D}\mathbf{K}^T(\mathbf{B} + \mathbf{K}\mathbf{D}\mathbf{K}^T)^{-1}(\mathbf{y} - \mathbf{K}\mathbf{d}). \tag{N.37}$$

Appendix O

DLM Conditional Independence

In this appendix, we show that $\mathbf{X}_i$ and $\mathbf{Y}_s$ for s such that $s > i$ are conditionally independent given $\mathbf{X}_{i+1}$. Meaning that

$$p(\mathbf{x}_i, \mathbf{y}_s|\mathbf{X}_{i+1}) = p(\mathbf{x}_i|\mathbf{X}_{i+1})p(\mathbf{y}_s|\mathbf{X}_{i+1}). \tag{O.1}$$

Since the random vectors $\mathbf{X}_i$ and $\mathbf{Y}_s$ are Gaussian distributed, proving that the covariance of every combination of elements of $\mathbf{X}_i$ and $\mathbf{Y}_s$ conditioned on $\mathbf{X}_{i+1}$, given as,

$$\text{Cov}[(\mathbf{X}_i)_j, (\mathbf{Y}_s)_k|\mathbf{X}_{i+1}] = \text{E}[(\mathbf{X}_i)_j(\mathbf{Y}_s)_k|\mathbf{X}_{i+1}] - \text{E}[(\mathbf{X}_i)_j|\mathbf{X}_{i+1}]\text{E}[(\mathbf{Y}_s)_k|\mathbf{X}_{i+1}], \tag{O.2}$$

is zero is sufficient to prove that they are independent (see Appendix W). In the above equation j and k represent the jth and kth elements of the random vectors $\mathbf{X}_i$ and $\mathbf{Y}_s$ respectively. So, if this condition is to be true it must be true for all combinations $j = 1, ..., n$ with $k = 1, ..., m$ where n and m are the lengths of the random vectors $\mathbf{X}_i$ and $\mathbf{Y}_s$ respectively.

$\mathbf{X}_i$ is independent of $\mathbf{X}_{i+1}$ and can be defined in terms of the $\mathbf{w}_i$ and $\mathbf{X}_0$ random vectors by Equation M.5 in Appendix M as,

$$\mathbf{X}_i = (\prod_{j=1}^{i} \mathbf{G}_j)\mathbf{X}_0 + \sum_{k=2}^{i+1}((\prod_{j=k}^{i} \mathbf{G}_j)\mathbf{w}_{k-1}). \tag{O.3}$$

So, since we typically choose $\text{E}[\mathbf{X}_0] = 0$ the expectation of $\mathbf{X}_i$ is zero, giving,

$$\text{E}[\mathbf{X}_i|\mathbf{X}_{i+1}] = \text{E}[\mathbf{X}_i] = 0. \tag{O.4}$$

This shows that the second term in Equation O.2 is 0 for all j and k combinations. What is left then is to show that

$$\text{E}[(\mathbf{X}_i)_j(\mathbf{Y}_s)_k|\mathbf{X}_{i+1}] = 0. \tag{O.5}$$

Starting with $s = i + 1$ we have,

$$\text{E}[(\mathbf{X}_i)_j(\mathbf{F}_{i+1}\mathbf{X}_{i+1} + \mathbf{v}_{i+1})_k|\mathbf{X}_{i+1}] = 0. \tag{O.6}$$

This is zero since $\mathbf{X}_0$ and $\mathbf{w}_s$ where $s \leq i$ (again, what the random vector $\mathbf{X}_i$ is composed of) have zero mean and are Gaussian distributed and are independent of $\mathbf{v}_{i+1}$ and the given $\mathbf{X}_{i+1}$. So, for all combinations of j and k this is zero. A similar story is seen to happen for $s = i + 2$, as,

$$\text{E}[(\mathbf{X}_i)_j(\mathbf{F}_{i+2}(\mathbf{F}_{i+1}\mathbf{X}_{i+1} + \mathbf{v}_{i+1}) + \mathbf{v}_{i+2})_k|\mathbf{X}_{i+1}] = 0, \tag{O.7}$$

and for any $s > i$. The breaking point is $s = i$ since for this and any smaller s, $\mathbf{Y}_s$ is no longer dependent on $\mathbf{X}_{i+1}$ and it can be seen that non-zero terms such as $\text{E}[(\mathbf{X}_0)_j(\mathbf{A}\mathbf{X}_0)_k]$

arise (where $\mathbf{A}$ is any matrix). So, the covariance $\text{Cov}[(\mathbf{X}_i)_j, (\mathbf{Y}_s)_k|\mathbf{X}_{i+1}]$ for s such that $s > i$ is zero for all combinations of j and k and therefore $\mathbf{X}_i$ and $\mathbf{Y}_s$ are conditionally independent given $\mathbf{X}_{i+1}$.

Appendix P

Matrix Calculus

In this thesis, we use the numerator-layout notation convention of matrix calculus rather than the denominator-layout convention.

Derivative of a Vector by a Scalar

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix} \tag{P.1}$$

Derivative of a Scalar by a Vector

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix} \tag{P.2}$$

Derivative of a Vector by a Vector

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \tag{P.3}$$

Derivative of a Scalar by a Matrix

$$\frac{\partial y}{\partial \mathbf{A}} = \begin{bmatrix} \frac{\partial y}{\partial A_{11}} & \frac{\partial y}{\partial A_{21}} & \cdots & \frac{\partial y}{\partial A_{m1}} \\ \frac{\partial y}{\partial A_{12}} & \frac{\partial y}{\partial A_{22}} & \cdots & \frac{\partial y}{\partial A_{m2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial A_{1n}} & \frac{\partial y}{\partial A_{2n}} & \cdots & \frac{\partial y}{\partial A_{mn}} \end{bmatrix} \tag{P.4}$$

Identities

All the following derivative identities can most easily be verified by explicitly writing out a low dimensional case, and then extending the work to N dimensions.

The scalar quantity $a = \mathbf{v}^T\mathbf{x} = \mathbf{x}^T\mathbf{v}$

$$\frac{\partial a}{\partial \mathbf{x}} = \mathbf{v}^T \tag{P.5}$$

or:

$$\left(\frac{\partial a}{\partial \mathbf{x}}\right)^T = \mathbf{v} \tag{P.6}$$

The scalar quantity $a = \mathbf{x}^T\mathbf{A}\mathbf{x}$

$$\frac{\partial a}{\partial \mathbf{x}} = \begin{cases} \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T) & \text{non sysmmetric } \mathbf{A} \\ 2\mathbf{x}^T\mathbf{A} & \text{sysmmetric } \mathbf{A} \end{cases} \tag{P.7}$$

or:

$$\left(\frac{\partial a}{\partial \mathbf{x}}\right)^T = \begin{cases} (\mathbf{A}^T + \mathbf{A})\mathbf{x} & \text{non sysmmetric } \mathbf{A} \\ 2\mathbf{A}\mathbf{x} & \text{sysmmetric } \mathbf{A} \end{cases} \tag{P.8}$$

The scalar quantity $a = (\mathbf{A}\mathbf{x} + \mathbf{b})^T\mathbf{C}(\mathbf{D}\mathbf{x} + \mathbf{e})$

$$\frac{\partial a}{\partial \mathbf{x}} = (\mathbf{D}\mathbf{x} + \mathbf{e})^T\mathbf{C}^T\mathbf{A} + (\mathbf{A}\mathbf{x} + \mathbf{b})^T\mathbf{C}\mathbf{D} \tag{P.9}$$

or:

$$\left(\frac{\partial a}{\partial \mathbf{x}}\right)^T = \mathbf{A}^T\mathbf{C}(\mathbf{D}\mathbf{x} + \mathbf{e}) + \mathbf{D}^T\mathbf{C}^T(\mathbf{A}\mathbf{x} + \mathbf{b}) \tag{P.10}$$

In the case where $\mathbf{C}$ is symmetric, $\mathbf{A} = \mathbf{D}$, and $\mathbf{b} = \mathbf{e}$, so that we have $a = (\mathbf{A}\mathbf{x}+\mathbf{b})^T\mathbf{C}(\mathbf{A}\mathbf{x}+\mathbf{b})$, the derivative is given by:

$$\frac{\partial a}{\partial \mathbf{x}} = 2(\mathbf{A}\mathbf{x} + \mathbf{b})^T\mathbf{C}\mathbf{A} \tag{P.11}$$

or:

$$\left(\frac{\partial a}{\partial \mathbf{x}}\right)^T = 2\mathbf{A}^T\mathbf{C}(\mathbf{A}\mathbf{x} + \mathbf{b}) \tag{P.12}$$

Furthermore, in the case where $\mathbf{C}$ is symmetric, $\mathbf{A} = \mathbf{D} = \mathbf{I}$, and $\mathbf{b} = \mathbf{e}$, so that we have $a = (\mathbf{x} + \mathbf{b})^T\mathbf{C}(\mathbf{x} + \mathbf{b})$, the derivative is given by:

$$\frac{\partial a}{\partial \mathbf{x}} = 2(\mathbf{x} + \mathbf{b})^T\mathbf{C} \tag{P.13}$$

or:

$$\left(\frac{\partial a}{\partial \mathbf{x}}\right)^T = 2\mathbf{C}(\mathbf{x} + \mathbf{b}) \tag{P.14}$$

The vector quantity $\mathbf{y} = \mathbf{A}\mathbf{x}$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A} \tag{P.15}$$

The vector quantity $\mathbf{y} = \mathbf{x}^T\mathbf{A}$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}^T \tag{P.16}$$

Consistent with these two identities is that:

$$\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^T = \frac{\partial \mathbf{y}^T}{\partial \mathbf{x}} \tag{P.17}$$

Trace of matrices

A square matrix AB

$$\frac{\partial \operatorname{tr}(\mathbf{A}\mathbf{B})}{\partial \mathbf{A}} = \frac{\partial \operatorname{tr}(\mathbf{B}^T\mathbf{A}^T)}{\partial \mathbf{A}} = \mathbf{B}^T \tag{P.18}$$

$\mathbf{A}\mathbf{B}\mathbf{A}^T$

$$\frac{\partial \operatorname{tr}(\mathbf{A}\mathbf{B}\mathbf{A}^T)}{\partial \mathbf{A}} = \begin{cases} \mathbf{A}(\mathbf{B} + \mathbf{B}^T) & \text{non sysmmetric } \mathbf{B} \\ 2\mathbf{A}\mathbf{B} & \text{sysmmetric } \mathbf{B} \end{cases} \tag{P.19}$$

Hessian Matrix

Suppose we have a scalar function $f(\mathbf{x})$ with a vectored input $\mathbf{x}$ of length n. The Hessian matrix of $f(\mathbf{x})$ is a n x n matrix of second order partial derivatives given as,

$$\mathbf{H}(f(\mathbf{x})) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}. \tag{P.20}$$

With the definition of derivatives of a scalar by a vector and a vector by a vector it is seen that the Hessian matrix can be written as,

$$\mathbf{H}(f(\mathbf{x})) = \left(\frac{\partial}{\partial \mathbf{x}}\left(\frac{\partial f}{\partial \mathbf{x}}\right)\right)^T = \frac{\partial}{\partial \mathbf{x}}\left(\frac{\partial f}{\partial \mathbf{x}}\right)^T. \tag{P.21}$$

Appendix Q

Equivalence of the DLM MMSE estimator and an Orthogonal Projection

In this appendix, we show that for Gaussian distributed $\mathbf{X}_i$ and $\mathbf{Y}_i$ (with zero mean) for the general DLM, the orthogonal projection of the elements of $\mathbf{X}_i$ onto the $\boldsymbol{\Upsilon}_n$ vector space is equal to the conditional expectation $\mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{Y}}_n]$, which is the DLM MMSE estimator.

We start by writing the elements of the random vector $\mathbf{X}_i$ equal to two components, the orthogonal projection of elements of $\mathbf{X}_i$ onto the $\boldsymbol{\Upsilon}_n$ space and the elements of $\mathbf{X}_i$ orthogonal to the $\boldsymbol{\Upsilon}_n$ space (i.e. the trick of Equation 4.138) as,

$$\mathbf{X}_i = \bar{\mathbf{X}}_i + \tilde{\mathbf{X}}_i. \tag{Q.1}$$

Since we are requiring that $\mathrm{E}[\mathbf{X}_i]$ and $\mathrm{E}[\mathbf{Y}_i]$ are zero for all i, we also have $\mathrm{E}[\bar{\mathbf{X}}_i] = 0$ since the elements of $\bar{\mathbf{X}}_i$ are linear combinations of the elements of the $\mathbf{Y}_i$'s, and therefore we must also have $\mathrm{E}[\tilde{\mathbf{X}}_i] = 0$. In the DLM formulation, having $\mathrm{E}[\mathbf{X}_i]$ and $\mathrm{E}[\mathbf{Y}_i]$ equal to zero only amounts to selecting the prior expectation $\mathrm{E}[\mathbf{X}_0]$ as zero. This can be seen easily from the expanded DLM evolution and observation equations in Appendix M.

Now, it can be seen that the elements of $\tilde{\mathbf{X}}_i$ are uncorrelated to every random variable that that is an element $\mathbf{Y}_s$ for any $s \leq n$. We see this since they have zero covariance, as,

$$\mathrm{E}[\tilde{X}_{ij} Y_{sk}^T] - \mathrm{E}[\tilde{X}_{ij}]\mathrm{E}[Y_{sk}]^T = 0, \tag{Q.2}$$

where we define $\tilde{X}_{ij}$ as the jth element of $\tilde{\mathbf{X}}_i$ and as Y_{sk} the kth element of $\mathbf{Y}_s$. The last term is zero for all j and k based on the discussion just had and the first term is clearly zero since these random variables are orthogonal to each other just by the way $\tilde{\mathbf{X}}_i$ was defined.

So, what we consider next, and this is where the Gaussian part comes into play, is that since Gaussian distributed random variables that are uncorrelated are also independent (see Appendix W) $\tilde{\mathbf{X}}_i$ and all the random variables contained in $\tilde{\mathbf{Y}}_n$ are independent (recall our notation $\tilde{\mathbf{Y}}_n = \mathbf{Y}_1, ..., \mathbf{Y}_n$). Therefore, we may write

$$\mathrm{E}[\tilde{\mathbf{X}}_i] = \mathrm{E}[\tilde{\mathbf{X}}_i|\tilde{\mathbf{Y}}_n] = 0, \tag{Q.3}$$

and from this we find that,

$$\begin{aligned}
\mathrm{E}[\tilde{\mathbf{X}}_i|\tilde{\mathbf{Y}}_n] &= \mathrm{E}[\mathbf{X}_i - \bar{\mathbf{X}}_i|\tilde{\mathbf{Y}}_n] && \text{(Q.4)}\\
&= \mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{Y}}_n] - \mathrm{E}[\bar{\mathbf{X}}_i|\tilde{\mathbf{Y}}_n] && \text{(Q.5)}\\
&= \mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{Y}}_n] - \bar{\mathbf{X}}_i && \text{(Q.6)}\\
&= 0. && \text{(Q.7)}
\end{aligned}$$

And so we have the orthogonal projection of the elements of $\mathbf{X}_i$ onto the $\boldsymbol{\Upsilon}_n$ vector space (i.e. $\bar{\mathbf{X}}_i$) equal to the conditional mean, as,

$$\bar{\mathbf{X}}_i = \mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{Y}}_n]. \tag{Q.8}$$

To summarize, we have shown that for Gaussian distributed $\mathbf{X}_i$ and $\mathbf{Y}_i$ with zero mean for the general DLM (which is in fact how we have defined the general DLM by choosing to specify $\mathbf{w}_i$, $\mathbf{v}_i$, and $\mathbf{X}_0$ as Gaussian distributed. The only caveat is we must choose $\mathrm{E}[\mathbf{X}_0]$ to be 0) the orthogonal projection of the elements of $\mathbf{X}_i$ onto the $\boldsymbol{\Upsilon}_n$ vector space is equal to the conditional expectation $\mathrm{E}[\mathbf{X}_i|\tilde{\mathbf{Y}}_n]$, which is the DLM MMSE estimator.

Orthogonality Principle

In addition to what was just proven, it will also be instructive to show that for any random variable X (this section has nothing to do with the DLM) the MMSE estimator of X restricted to be on a given vector space is the orthogonal projection of the random variable onto that vector space. This is known as the orthogonality principle.

Proof

We start by writing the random variable X in two different ways. First, using the trick of Equation 4.136, we write,

$$X = \bar{X} + \tilde{X}, \tag{Q.9}$$

where $\bar{X}$ is the orthogonal projection onto any vector space and $\tilde{X}$ is orthogonal to this vector space. Second, we write,

$$X = \bar{W} + W', \tag{Q.10}$$

where $\bar{W}$ is any random vector in the same vector space as $\bar{X}$ and W' is the random variable that is needed to be added to $\bar{W}$ to give X (note that W' is only orthogonal to the vector space in the one case where $\bar{W} = \bar{X}$).

It will be shown now that $\mathrm{E}[(X - \bar{W})^2] \geq \mathrm{E}[(X - \bar{X})^2]$. This inequality states that the orthogonal projection onto the vector space $\bar{X}$ has the smallest MSE out of any other possible estimator on the vector space, hence completing our proof. We have,

$$\begin{aligned} \mathrm{E}[(X - \bar{W})^2] &= \mathrm{E}[(X + \bar{X} - \bar{X} - \bar{W})^2] && \text{(Q.11)} \\ &= \mathrm{E}[(X - \bar{X})^2] + \mathrm{E}[(\bar{X} - \bar{W})^2] + 2\mathrm{E}[(X - \bar{X})(\bar{X} - \bar{W})]. && \text{(Q.12)} \end{aligned}$$

The last term is zero since $X - \bar{X}$ is orthogonal to the given vector space and $\bar{X} - \bar{W}$ is in the given vector space. Therefore we see that

$$\mathrm{E}[(X - \bar{W})^2] \geq \mathrm{E}[(X - \bar{X})^2], \tag{Q.13}$$

since $\mathrm{E}[(\bar{X} - \bar{W})^2]$ is positive, and the proof is complete.

As an aside, note that this orthogonality principle is fundamentally similar to a result from linear algebra that says the square difference between a vector and some vector on a vector space is minimized when the vector on the vector space is the orthogonal projection of the other vector. If we call the other vector x, the vector on the vector space $\bar{w}$, and the orthogonal projection of x onto the vector space $\bar{x}$, then we have,

$$(x - \bar{w})^2 \geq (x - \bar{x})^2, \tag{Q.14}$$

and the proof is fundamentally the same as what we have shown here for random variables rather than vectors. And this notion here can be thought of as just reflecting basic geometry.

Also, for the extension of the orthogonality principal to random vectors, the proof that $\mathrm{E}[(\mathbf{X} - \bar{\mathbf{W}})^T(\mathbf{X} - \bar{\mathbf{W}})] \geq \mathrm{E}[(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}})]$ is essentially the same. If we define

$$\mathbf{X} = \bar{\mathbf{X}} + \tilde{\mathbf{X}}, \tag{Q.15}$$

and

$$\mathbf{X} = \bar{\mathbf{W}} + \mathbf{W}', \tag{Q.16}$$

and carry out the same steps, we have,

$$\begin{aligned}
\mathrm{E}[(\mathbf{X} - \bar{\mathbf{W}})^T(\mathbf{X} - \bar{\mathbf{W}})] &= \mathrm{E}[(\mathbf{X} + \bar{\mathbf{X}} - \bar{\mathbf{X}} - \bar{\mathbf{W}})^T(\mathbf{X} + \bar{\mathbf{X}} - \bar{\mathbf{X}} - \bar{\mathbf{W}})] && \text{(Q.17)} \\
&= \mathrm{E}[(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}})] + \mathrm{E}[(\bar{\mathbf{X}} - \bar{\mathbf{W}})^T(\bar{\mathbf{X}} - \bar{\mathbf{W}})] \\
&\quad + \mathrm{E}[(\mathbf{X} - \bar{\mathbf{X}})^T(\bar{\mathbf{X}} - \bar{\mathbf{W}})] + \mathrm{E}[(\bar{\mathbf{X}} - \bar{\mathbf{W}})^T(\mathbf{X} - \bar{\mathbf{X}})] && \text{(Q.18)} \\
&= \mathrm{E}[(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}})] + \mathrm{E}[(\bar{\mathbf{X}} - \bar{\mathbf{W}})^T(\bar{\mathbf{X}} - \bar{\mathbf{W}})]. && \text{(Q.19)}
\end{aligned}$$

Therefore we have,

$$\mathrm{E}[(\mathbf{X} - \bar{\mathbf{W}})^T(\mathbf{X} - \bar{\mathbf{W}})] \geq \mathrm{E}[(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}})] \tag{Q.20}$$

since the vector $\mathrm{E}[(\bar{\mathbf{X}} - \bar{\mathbf{W}})^T(\bar{\mathbf{X}} - \bar{\mathbf{W}})]$ has all positive elements, and the proof for the multivariate random vector case is complete.

Appendix R

Proofs to Three Results

In this appendix, the three results presented in Section 4.5.5.2 are proven.

Result 1

We have the function,

$$F(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{a})^T\mathbf{A}(\mathbf{x}-\mathbf{a}) + \frac{1}{2}(\mathbf{b}-\mathbf{B}\mathbf{x})^T\mathbf{C}(\mathbf{b}-\mathbf{B}\mathbf{x}), \tag{R.1}$$

where $\mathbf{A}$ and $\mathbf{C}$ are symmetric matrices. Setting its derivative equal to zero gives,

$$\left(\frac{\partial F(\mathbf{x})}{\partial \mathbf{x}}\right)^T = \mathbf{A}(\mathbf{x}_m-\mathbf{a}) - \mathbf{B}^T\mathbf{C}(\mathbf{b}-\mathbf{B}\mathbf{x}_m) = 0. \tag{R.2}$$

So we have,

$$\mathbf{x}_m = (\mathbf{A}+\mathbf{B}^T\mathbf{C}\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{C}\mathbf{b}+\mathbf{A}\mathbf{a}), \tag{R.3}$$

which can also be written as (see Equations N.26 through N.37 in Appendix N),

$$\mathbf{x}_m = \mathbf{a} + \mathbf{A}^{-1}\mathbf{B}^T(\mathbf{C}^{-1}+\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)^{-1}(\mathbf{b}-\mathbf{B}\mathbf{a}). \tag{R.4}$$

Result 2

We have the function,

$$F(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{a})^T\mathbf{A}(\mathbf{x}-\mathbf{a}) + \frac{1}{2}(\mathbf{b}-\mathbf{B}\mathbf{x})^T\mathbf{C}(\mathbf{b}-\mathbf{B}\mathbf{x}) + \frac{1}{2}(\mathbf{d}-\mathbf{D}\mathbf{x})^T\mathbf{E}(\mathbf{d}-\mathbf{D}\mathbf{x}), \tag{R.5}$$

where $\mathbf{A}$, $\mathbf{C}$, and $\mathbf{E}$ are symmetric matrices. Setting its derivative equal to zero gives,

$$\left(\frac{\partial F(\mathbf{x})}{\partial \mathbf{x}}\right)^T = \mathbf{A}(\mathbf{x}_m-\mathbf{a}) - \mathbf{B}^T\mathbf{C}(\mathbf{b}-\mathbf{B}\mathbf{x}_m) - \mathbf{D}^T\mathbf{E}(\mathbf{d}-\mathbf{D}\mathbf{x}_m) = 0. \tag{R.6}$$

So, if we make the substitutions $\mathbf{Z} = \mathbf{A}+\mathbf{B}^T\mathbf{C}\mathbf{B}$ and $\mathbf{Y} = \mathbf{B}^T\mathbf{C}\mathbf{b}+\mathbf{A}\mathbf{a}$ we have,

$$\mathbf{x}_m = (\mathbf{Z}+\mathbf{D}^T\mathbf{E}\mathbf{D})^{-1}(Y+\mathbf{D}^T\mathbf{E}\mathbf{d}). \tag{R.7}$$

Now, some algebra will be done to put this into the desired form that is given in Section 4.5.5.2. Using the Woodbury matrix identity of Appendix I we may write,

$$(\mathbf{Z}+\mathbf{D}^T\mathbf{E}\mathbf{D})^{-1}=\mathbf{Z}^{-1}-\mathbf{Z}^{-1}\mathbf{D}^T(\mathbf{E}^{-1}+\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{Z}^{-1}. \tag{R.8}$$

Therefore $\mathbf{x}_m$ can be written as,

$$\begin{aligned}\mathbf{x}_m=&\mathbf{Z}^{-1}Y-\mathbf{Z}^{-1}\mathbf{D}^T(\mathbf{E}^{-1}+\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{Z}^{-1}Y+\mathbf{Z}^{-1}\mathbf{D}^T\mathbf{E}\mathbf{d}\\&-\mathbf{Z}^{-1}\mathbf{D}^T(\mathbf{E}^{-1}+\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T\mathbf{E}\mathbf{d},\end{aligned} \tag{R.9}$$

where last two terms are equal to,

$$\mathbf{Z}^{-1}\mathbf{D}^T(\mathbf{E}-(\mathbf{E}^{-1}+\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T\mathbf{E})\mathbf{d}. \tag{R.10}$$

Using the matrix addition inversion identity of Appendix I we may write,

$$\mathbf{E}-(\mathbf{E}^{-1}+\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T\mathbf{E}=(\mathbf{E}^{-1}+\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T)^{-1}. \tag{R.11}$$

So, $\mathbf{x}_m$ can be written as,

$$\begin{aligned}\mathbf{x}_m=&\mathbf{Z}^{-1}Y-\mathbf{Z}^{-1}\mathbf{D}^T(\mathbf{E}^{-1}+\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{Z}^{-1}Y\\&+\mathbf{Z}^{-1}\mathbf{D}^T(\mathbf{E}^{-1}+\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T)^{-1}\mathbf{d} &\text{(R.12)}\\=&\mathbf{Z}^{-1}Y+\mathbf{Z}^{-1}\mathbf{D}^T(\mathbf{E}^{-1}+\mathbf{D}\mathbf{Z}^{-1}\mathbf{D}^T)^{-1}(\mathbf{d}-\mathbf{D}\mathbf{Z}^{-1}Y). &\text{(R.13)}\end{aligned}$$

Result 3

We have the function,

$$F(\mathbf{x}_1,\mathbf{x}_2)=k+\frac{1}{2}(\mathbf{x}_1-\mathbf{a})\mathbf{A}(\mathbf{x}_1-\mathbf{a})+\frac{1}{2}(\mathbf{b}-\mathbf{B}\mathbf{x}_2)^T\mathbf{C}(\mathbf{b}-\mathbf{B}\mathbf{x}_2)+\frac{1}{2}(\mathbf{x}_2-\mathbf{D}\mathbf{x}_1)^T\mathbf{E}(\mathbf{x}_2-\mathbf{D}\mathbf{x}_1), \tag{R.14}$$

where $\mathbf{A}$ and $\mathbf{E}$ are symmetric matrices. We define $\mathbf{w}=\mathbf{x}_2-\mathbf{D}\mathbf{x}_1$ so that we can write,

$$F(\mathbf{w},\mathbf{x}_2)=k+\frac{1}{2}(\mathbf{D}^{-1}(\mathbf{x}_2-\mathbf{w})-\mathbf{a})\mathbf{A}(\mathbf{D}^{-1}(\mathbf{x}_2-\mathbf{w})-\mathbf{a})+\frac{1}{2}(\mathbf{b}-\mathbf{B}\mathbf{x}_2)^T\mathbf{C}(\mathbf{b}-\mathbf{B}\mathbf{x}_2)+\frac{1}{2}\mathbf{w}^T\mathbf{E}\mathbf{w}. \tag{R.15}$$

Setting the partial derivative with respect to $\mathbf{w}$ equal to zero we have,

$$\left(\frac{\partial F(\mathbf{w},\mathbf{x}_2)}{\partial\mathbf{w}}\right)^T=-\mathbf{D}^{-1^T}\mathbf{A}(\mathbf{D}^{-1}(\mathbf{x}_2-\mathbf{w}_m)-\mathbf{a})-\mathbf{E}\mathbf{w}_m=0. \tag{R.16}$$

Therefore for $\mathbf{w}_m$ we have,

$$\mathbf{w}_m=(\mathbf{D}^{-1^T}\mathbf{A}\mathbf{D}^{-1}-\mathbf{E})^{-1}\mathbf{D}^{-1^T}\mathbf{A}(\mathbf{D}^{-1}\mathbf{x}_2-\mathbf{a}). \tag{R.17}$$

Using the Woodbury matrix identity of Appendix I we may write,

$$(\mathbf{D}^{-1^T}\mathbf{A}\mathbf{D}^{-1} - \mathbf{E})^{-1} = \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T - \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T(\mathbf{E}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T. \tag{R.18}$$

So we have,

$$\begin{aligned}\mathbf{w}_m &= (\mathbf{D} - \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T(\mathbf{E}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T)^{-1}\mathbf{D})(\mathbf{D}^{-1}\mathbf{x}_2 - \mathbf{a}) &\text{(R.19)}\\ &= (\mathbf{I} - \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T(\mathbf{E}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T)^{-1})(\mathbf{x}_2 - \mathbf{D}\mathbf{a}). &\text{(R.20)}\end{aligned}$$

We define $\mathbf{P} = \mathbf{E}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T$ (note that since $\mathbf{E}$ and $\mathbf{A}$ are symmetric $\mathbf{P}$ must be symmetric) and $\bar{\mathbf{w}} = \mathbf{x}_2 - \mathbf{D}\mathbf{a}$ so that we can write,

$$\begin{aligned}\mathbf{w}_m &= (\mathbf{I} - \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T\mathbf{P}^{-1})\bar{\mathbf{w}} &\text{(R.21)}\\ &= (\mathbf{P} - \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T)\mathbf{P}^{-1}\bar{\mathbf{w}} &\text{(R.22)}\\ &= \mathbf{E}^{-1}\mathbf{P}^{-1}\bar{\mathbf{w}}. &\text{(R.23)}\end{aligned}$$

We note that we can write $\mathbf{D}^{-1}(\mathbf{x}_2 - \mathbf{w}) - \mathbf{a}$ as,

$$\begin{aligned}\mathbf{D}^{-1}(\mathbf{x}_2 - \mathbf{w}) - \mathbf{a} &= \mathbf{D}^{-1}(\mathbf{x}_2 - \mathbf{w} - \mathbf{D}\mathbf{a}) &\text{(R.24)}\\ &= \mathbf{D}^{-1}(\bar{\mathbf{w}} - \mathbf{w}). &\text{(R.25)}\end{aligned}$$

So the first term of $F(\mathbf{x}_1, \mathbf{x}_2)$ can be written as,

$$\frac{1}{2}(\mathbf{D}^{-1}(\mathbf{x}_2 - \mathbf{w}) - \mathbf{a})\mathbf{A}(\mathbf{D}^{-1}(\mathbf{x}_2 - \mathbf{w}) - \mathbf{a}) = \frac{1}{2}(\bar{\mathbf{w}} - \mathbf{w})^T\mathbf{D}^{-1^T}\mathbf{A}\mathbf{D}^{-1}(\bar{\mathbf{w}} - \mathbf{w}). \tag{R.26}$$

Plugging in $\mathbf{w}_m$ in the form of Equation R.21 for $\mathbf{w}$ into $F(\mathbf{w}, \mathbf{x}_2)$ we have,

$$\begin{aligned}F(\mathbf{w}_m, \mathbf{x}_2) = k &+ \frac{1}{2}(\bar{\mathbf{w}} - (\mathbf{I} - \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T\mathbf{P}^{-1})\bar{\mathbf{w}})^T\mathbf{D}^{-1^T}\mathbf{A}\mathbf{D}^{-1}(\bar{\mathbf{w}} - (\mathbf{I} - \mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T\mathbf{P}^{-1})\bar{\mathbf{w}})\\ &+ \frac{1}{2}(\mathbf{b} - \mathbf{B}\mathbf{x}_2)^T\mathbf{C}(\mathbf{b} - \mathbf{B}\mathbf{x}_2) + \frac{1}{2}(\mathbf{E}^{-1}\mathbf{P}^{-1}\bar{\mathbf{w}})^T\mathbf{E}(\mathbf{E}^{-1}\mathbf{P}^{-1}\bar{\mathbf{w}}). \qquad \text{(R.27)}\end{aligned}$$

This is the same as function we have called $F(\mathbf{x}_{1m}(\mathbf{x}_2), \mathbf{x}_2)$ in Section 4.5.5.2, because $\mathbf{w}_m$ is a function of $\mathbf{x}_2$ as well. The job now is to simplify this to be in the desired form that we gave in Section 4.5.5.2. The last term is given as,

$$\text{last term} = \frac{1}{2}\bar{\mathbf{w}}^T\mathbf{P}^{-1}\mathbf{E}^{-1}\mathbf{P}^{-1}\bar{\mathbf{w}}. \tag{R.28}$$

The first non constant term is given as,

$$\text{first term} = \frac{1}{2}(\mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T\mathbf{P}^{-1}\bar{\mathbf{w}})^T\mathbf{D}^{-1^T}\mathbf{A}\mathbf{D}^{-1}(\mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T\mathbf{P}^{-1}\bar{\mathbf{w}}) \tag{R.29}$$

$$= \frac{1}{2}(\mathbf{D}\mathbf{A}^{-1}\mathbf{D}^T\mathbf{P}^{-1}\bar{\mathbf{w}})^T\mathbf{P}^{-1}\bar{\mathbf{w}} \tag{R.30}$$

$$= \frac{1}{2}\bar{\mathbf{w}}^T\mathbf{P}^{-1}\mathbf{D}\mathbf{A}^{-1^T}\mathbf{D}^T\mathbf{P}^{-1}\bar{\mathbf{w}} \tag{R.31}$$

$$= \frac{1}{2}\bar{\mathbf{w}}^T\mathbf{P}^{-1}(\mathbf{P} - \mathbf{E}^{-1})\mathbf{P}^{-1}\bar{\mathbf{w}}. \tag{R.32}$$

So we have,

$$\text{first term + last term} = \frac{1}{2}\bar{\mathbf{w}}^T\mathbf{P}^{-1}\mathbf{P}\mathbf{P}^{-1}\bar{\mathbf{w}} \tag{R.33}$$

$$= \frac{1}{2}\bar{\mathbf{w}}^T\mathbf{P}^{-1}\bar{\mathbf{w}}. \tag{R.34}$$

Therefore, the function $F(\mathbf{x}_{1m}(\mathbf{x}_2), \mathbf{x}_2)$ is given as,

$$F(\mathbf{x}_{1m}, \mathbf{x}_2) = k + \frac{1}{2}(\mathbf{b} - \mathbf{B}\mathbf{x}_2)^T\mathbf{C}(\mathbf{b} - \mathbf{B}\mathbf{x}_2) + \frac{1}{2}(\mathbf{x}_2 - \mathbf{D}\mathbf{a})^T\mathbf{P}^{-1}(\mathbf{x}_2 - \mathbf{D}\mathbf{a}). \tag{R.35}$$

Appendix S

Multivariate Taylor Series

The Taylor series quadratic approximation for a univariate function $f(x)$ about the point a is given as,

$$f(x) = f(a) + \frac{df(x)}{dx}\Big|_{x=a}(x-a) + \frac{1}{2}\frac{d^2f(x)}{dx^2}\Big|_{x=a}(x-a)^2. \tag{S.1}$$

Similarly, the Taylor series quadratic approximation of a multivariate function $f(\mathbf{x})$ about the point $\mathbf{a}$ is given as,

$$f(\mathbf{x}) = f(\mathbf{a}) + \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{a}}(\mathbf{x}-\mathbf{a}) + \frac{1}{2}(\mathbf{x}-\mathbf{a})^T H(f(\mathbf{x}))\big|_{(\mathbf{x}-\mathbf{a})}(\mathbf{x}-\mathbf{a}). \tag{S.2}$$

This equation makes use of the matrix calculus defined in Appendix P. With this matrix calculus we note that the Hessian matrix $H(f(\mathbf{x}))$ is also the same as $\left(\frac{\partial}{\partial \mathbf{x}}\left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)\right)^T$ or $\frac{\partial}{\partial \mathbf{x}}\left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)^T$ as we have defined it. Also, note that the second term can alternatively be written as $(\mathbf{x}-\mathbf{a})^T(\frac{df(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{a}})^T$.

Appendix T

Multiple Regression DLM Estimation Reducing to MLR GLS Estimation

In this appendix, we show that for the multiple regression DLM with $\mathbf{W_i} = \mathbf{0}$, the filtering recurrence relations reduce to the weighted RLS recurrence relations and the smoothing recurrence relations reduce to $\mathbf{x}_i^n = \mathbf{x}_{i+1}^n$ and $\mathbf{P}_i^n = \mathbf{P}_{i+1}^n$. This effectively creates the same results as an MLR model estimation, as discussed in Section 4.6.2.

First, note that the one-step-ahead predictions $\mathbf{x}_i^{i-1}$ and $\mathbf{P}_i^{i-1}$ are

$$\begin{aligned} \mathbf{x}_i^{i-1} &= \mathbf{G}_i \mathbf{x}_{i-1}^{i-1} && \text{(T.1)} \\ &= \mathbf{x}_{i-1}^{i-1}, && \text{(T.2)} \end{aligned}$$

and

$$\begin{aligned} \mathbf{P}_i^{i-1} &= \mathbf{G}_i \mathbf{P}_{i-1}^{i-1} \mathbf{G}_i^T + \mathbf{W}_i && \text{(T.3)} \\ &= \mathbf{P}_{i-1}^{i-1}, && \text{(T.4)} \end{aligned}$$

and that the Kalman gain matrix $\mathbf{K}_i$ becomes

$$\begin{aligned} \mathbf{K}_i &= \mathbf{P}_i^{i-1} \mathbf{F}_i^T (\mathbf{V}_i + \mathbf{F}_i \mathbf{P}_i^{i-1} \mathbf{F}_i^T)^{-1} && \text{(T.5)} \\ &= \frac{\mathbf{P}_{i-1}^{i-1} \boldsymbol{\varphi}_i}{V_i + \boldsymbol{\varphi}_i^T \mathbf{P}_{i-1}^{i-1} \boldsymbol{\varphi}_i}. && \text{(T.6)} \end{aligned}$$

The estimates $\mathbf{x}_i^i$ and $\mathbf{P}_i^i$ are then seen to be,

$$\begin{aligned} \mathbf{x}_i^i &= \mathbf{x}_i^{i-1} + \mathbf{K}_i (y_i - \mathbf{F}_i \mathbf{x}_i^{i-1}) && \text{(T.7)} \\ \mathbf{x}_i^i &= \mathbf{x}_{i-1}^{i-1} + \frac{\mathbf{P}_{i-1}^{i-1} \boldsymbol{\varphi}_i}{V_i + \boldsymbol{\varphi}_i^T \mathbf{P}_{i-1}^{i-1} \boldsymbol{\varphi}_i} (y_i - \boldsymbol{\varphi}_i^T \mathbf{x}_{i-1}^{i-1}), && \text{(T.8)} \end{aligned}$$

and

$$\begin{aligned} \mathbf{P}_i^i &= \mathbf{P}_i^{i-1} - \mathbf{K}_i \mathbf{F}_i \mathbf{P}_i^{i-1} && \text{(T.9)} \\ \mathbf{P}_i^i &= \mathbf{P}_{i-1}^{i-1} - \frac{\mathbf{P}_{i-1}^{i-1} \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T \mathbf{P}_{i-1}^{i-1}}{V_i + \boldsymbol{\varphi}_i^T \mathbf{P}_{i-1}^{i-1} \boldsymbol{\varphi}_i}. && \text{(T.10)} \end{aligned}$$

We see that these recurrence relations are identical to the weighted RLS recurrence relations given in Appendix L, with the multiple regression DLMs V_i's defining the covariance structure of the errors.

Note that for the weighted RLS algorithm in Appendix L, the definition of $\mathbf{P}$ is $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$. So, we make the point that this is the same as the GLS covariance matrix of the regression coefficient estimator, $a^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$. The factor a^2 is just simply not factored out of the covariance structure in the DLM and RLS formulations like it is for the GLS MLR formulation.

Lastly, we will show that the smoothing recurrence relations for this DLM result in $\mathbf{x}_i^n = \mathbf{x}_{i+1}^n$ and $\mathbf{P}_i^n = \mathbf{P}_{i+1}^n$. In words, this means that all the smoothed estimates $\mathbf{x}_i^n$ and $\mathbf{P}_i^n$ are equal to the final filtered estimates $\mathbf{x}_n^n$ and $\mathbf{P}_n^n$, which as we just proved, the result of the weighted RLS algorithm, equal to the GLS estimation so long as the prior information is negligible.

Recall that for this DLM that we have $\mathbf{P}_{i+1}^i = \mathbf{P}_i^i$. Therefore, the matrix J_i in the smoothing recurrence relations becomes the identity matrix as follows:

$$\begin{aligned}
\mathbf{J}_i &= \mathbf{P}_i^i \mathbf{G}_{i+1}^T {\mathbf{P}_{i+1}^i}^{-1} && \text{(T.11)}\\
&= \mathbf{P}_i^i {\mathbf{P}_i^i}^{-1} && \text{(T.12)}\\
&= \mathbf{I}. && \text{(T.13)}
\end{aligned}$$

Similarly, recall that we have $\mathbf{x}_{i+1}^i = \mathbf{x}_i^i$. Therefore we see that the smoothing recurrence relations become

$$\begin{aligned}
\mathbf{x}_i^n &= \mathbf{x}_i^i + \mathbf{J}_i(\mathbf{x}_{i+1}^n - \mathbf{x}_{i+1}^i) && \text{(T.14)}\\
&= \mathbf{x}_i^i + (\mathbf{x}_{i+1}^n - \mathbf{x}_i^i) && \text{(T.15)}\\
&= \mathbf{x}_{i+1}^n, && \text{(T.16)}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{P}_i^n &= \mathbf{P}_i^i + \mathbf{J}_i(\mathbf{P}_{i+1}^n - \mathbf{P}_{i+1}^i)\mathbf{J}_i^T && \text{(T.17)}\\
&= \mathbf{P}_i^i + (\mathbf{P}_{i+1}^n - \mathbf{P}_i^i) && \text{(T.18)}\\
&= \mathbf{P}_{i+1}^n. && \text{(T.19)}
\end{aligned}$$

Appendix U

Local Level Trend DLM Forward Difference

In this appendix, we show that for the local level trend DLM with model matrix,

$$\mathbf{W}_i = \begin{bmatrix} 0 & 0 \\ 0 & \sigma^2_{trend} \end{bmatrix}, \tag{U.1}$$

the trend component of the sequence of smoothed estimates is equal to the forward difference of the level component of the sequence of smoothed estimates. Furthermore, it is shown that this is not true for when the model matrix is

$$\mathbf{W}_i = \begin{bmatrix} W_1 & 0 \\ 0 & W_2 \end{bmatrix}. \tag{U.2}$$

Proof

We define the smoothed estimate as,

$$\mathbf{x}_i^n = \begin{bmatrix} \mu_i^n \\ \alpha_i^n \end{bmatrix}. \tag{U.3}$$

Then, it must be shown that the smoothing recurrence relation,

$$\mathbf{x}_i^n = \mathbf{x}_i^i + \mathbf{P}_i^i \mathbf{G}_{i+1}^T {\mathbf{P}_{i+1}^i}^{-1} (\mathbf{x}_{i+1}^n - \mathbf{G}_{i+1}\mathbf{x}_i^i), \tag{U.4}$$

using the defined model matrices of this DLM and any filtered estimate x_i^i and covariance matrix P_i^i, is consistent with the forward difference equation,

$$\alpha_i^n = \mu_{i+1}^n - \mu_i^n. \tag{U.5}$$

The algebra for this becomes fairly involved. Following the Mathematica notebook below we see that it shows that we indeed have

$$\alpha_i^n + \mu_i^n = \mu_{i+1}^n. \tag{U.6}$$

Lastly, the second block of the Mathematica notebook shows that when we do not choose the first element of $\mathbf{W}_i$ to be zero, the forward difference no longer holds.

With zero variance on the background level evolution

In[204]:=
```
G_{i+1} = {{1, 1}, {1, 0}};
x_{i,i} = {μ_{i,i}, α_{i,i}};
W_i = {{0, 0}, {0, σ²_trend}};
P_{i,i} = {{a, b}, {b, c}};
P_{i+1,i} = G_{i+1}.P_{i,i}.Transpose[G_{i+1}] + W_i;
x_{i+1,n} = {μ_{i+1,n}, α_{i+1,n}};
x_{i,n} = x_{i,i} + P_{i,i}.Transpose[G_{i+1}].Inverse[P_{i+1,i}].(x_{i+1,n} - G_{i+1}.x_{i,i});
FullSimplify[x_{i,n}[[1]] + x_{i,n}[[2]]]
```

Out[211]= $\mu_{1+i,n}$

With variance on both the background level and trend evolution

In[212]:=
```
W_i = {{d, 0}, {0, e}};
P_{i+1,i} = G_{i+1}.P_{i,i}.Transpose[G_{i+1}] + W_i;
x_{i,n} = x_{i,i} + P_{i,i}.Transpose[G_{i+1}].Inverse[P_{i+1,i}].(x_{i+1,n} - G_{i+1}.x_{i,i});
FullSimplify[x_{i,n}[[1]] + x_{i,n}[[2]]]
```

Out[215]=
$$-\frac{d\,(a+e)\,\alpha_{i,i} + (a+b)\,d\,\alpha_{1+i,n} + d\,(-b+e)\,\mu_{i,i} + (-b^2 + a\,c + (a+2\,b+c)\,e)\,\mu_{1+i,n}}{b^2 - (2\,b+c+d)\,e - a\,(c+d+e)}$$

Appendix V

MCMC Results on the SOO Data Record

To illustrate the MCMC samples of the unknown DLM parameters, and to also assess the convergence of the MCMC chain, we show histograms and trace plots in Figures V.1 through V.12 like what was shown in Figure 6.2 for the single time series example. In these figures, the histograms and trace plots for each unknown parameter σ_{trend}, σ_{AR}, and ρ are shown on large grids of altitude and latitude. We see from the trace plots that the chains for each altitude-latitude region for the SOO data record are fairly well converged.

A few other points from these illustrated MCMC samples can be made. We see that the values of σ_{trend} do not reach much higher than 0.00025 for all of the altitude-latitude regions. Recall that σ_{trend} indicates the degree of variation for the background level fit. If we recall the example in Section 5.1.2 with Figure 6.18, we can roughly infer that our background level fits are at most varying somewhere between what is shown in (a) and (b) in the figure and typically are closer to (a) or even lesser than (a). The only exceptions to this are a few altitude-latitude regions in the tropical lower altitudes.

For an easier way to look at the samples of σ_{trend}, we make heat maps with dimensions of altitude and latitude for the mean, 5th percentile, and 95th percentile of the samples. This is shown in Figure V.13. Similarly, we show this same concept for the other two parameters. These are shown in Figures V.14 and V.15.

For Figure V.15, which illustrates the samples of ρ, except for the altitude-latitude regions in the lower and mid tropical regions, we see that the data does not favour any particular value of ρ. In more detail, from the histograms for the samples (Figures V.9 and V.11), we see that the marginal distributions are all fairly constant over 0 to 1. However, for the lower and mid tropical regions the story is different. In these regions, the model is telling us with a decent amount of certainty that the autocorrelation is high, especially from 30.5 to 35.5 km. Figure V.15 is directly comparable to the Prais-Winsten illustration in Figure 2.13. We re-show Figure 2.13 in Figure V.16 for comparison. We see that the results of these are quite different.

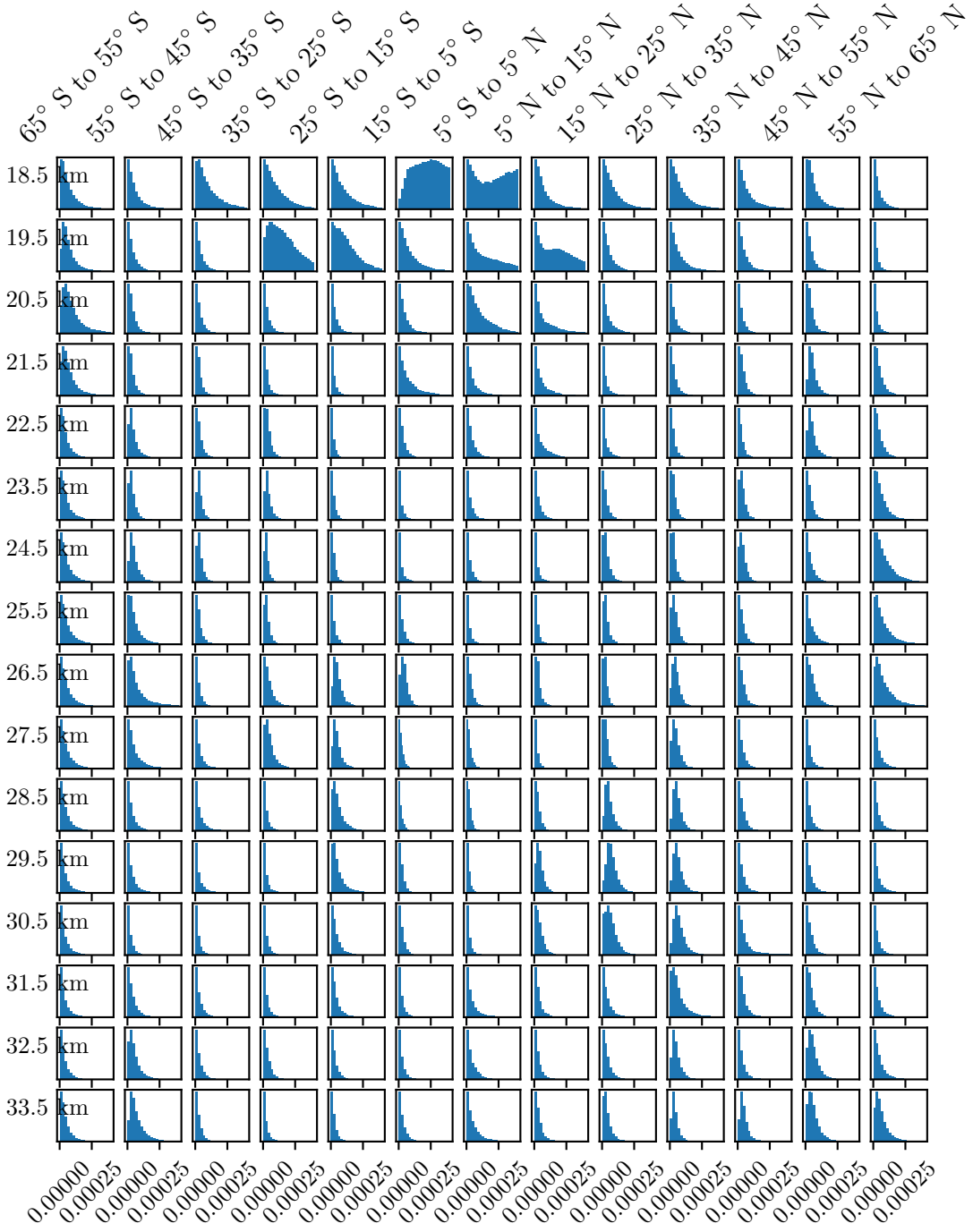


Figure V.1: σ_{trend} Histograms.

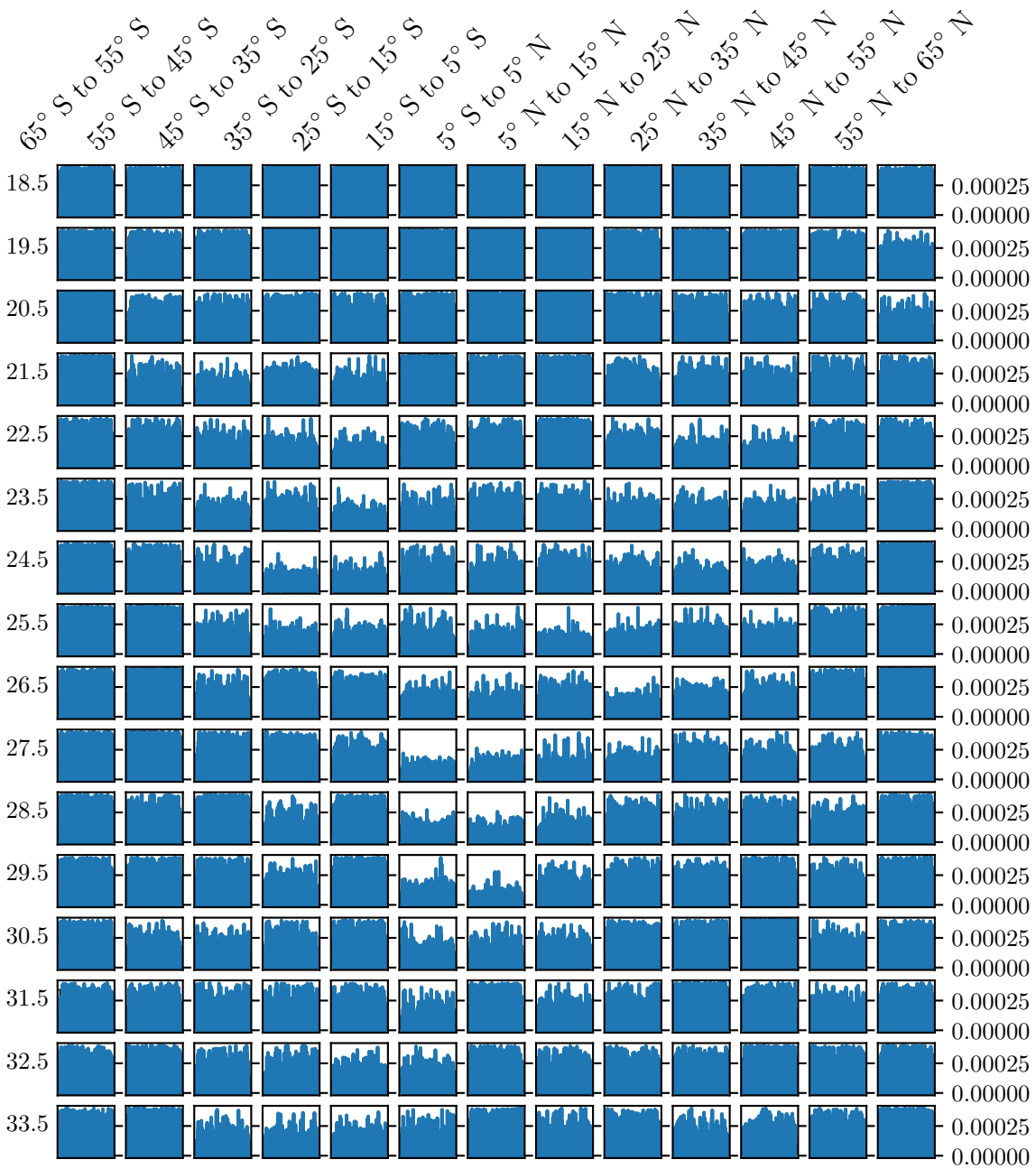


Figure V.2: σ_{trend} Trace Plots.

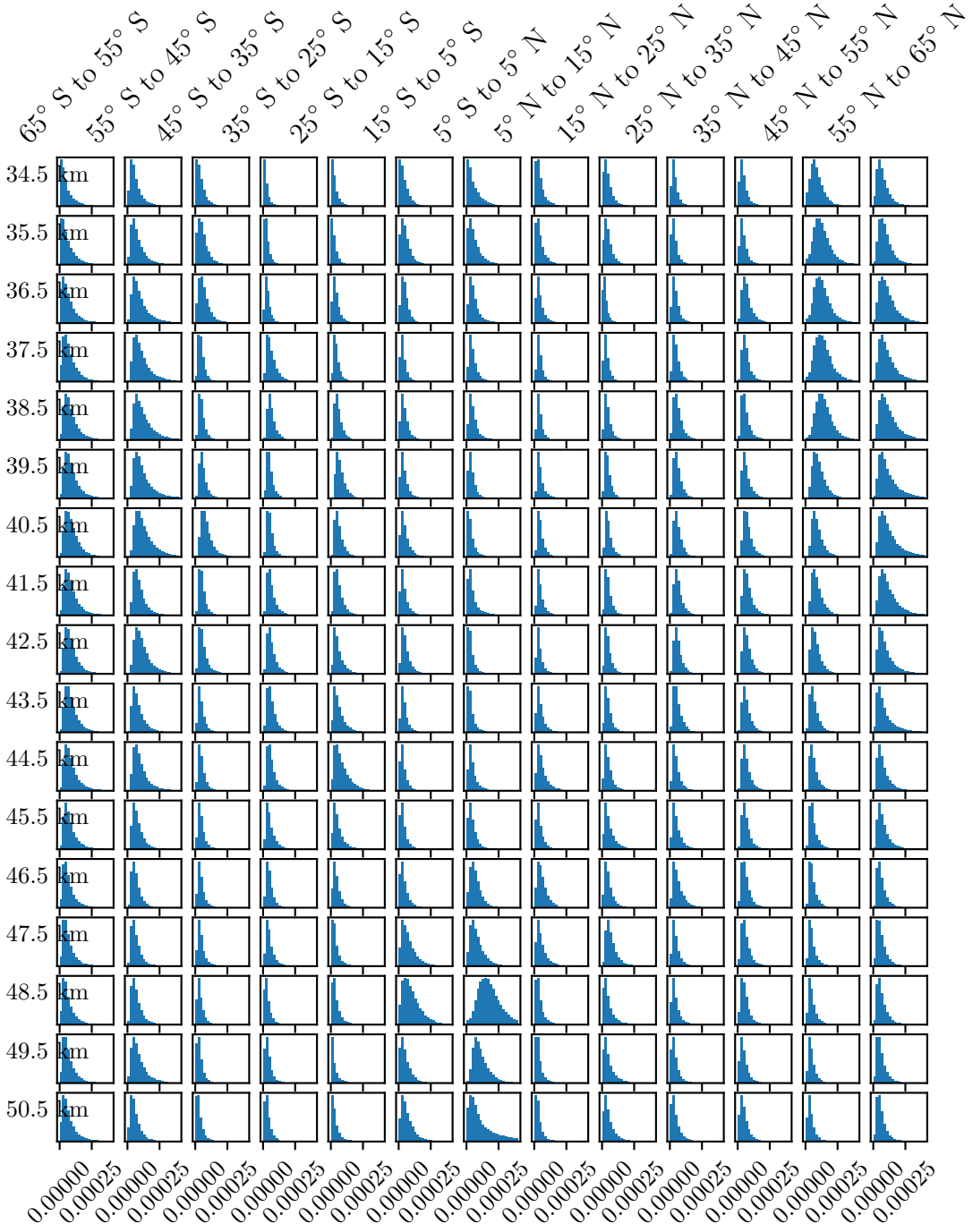


Figure V.3: σ_{trend} Histograms Continued.

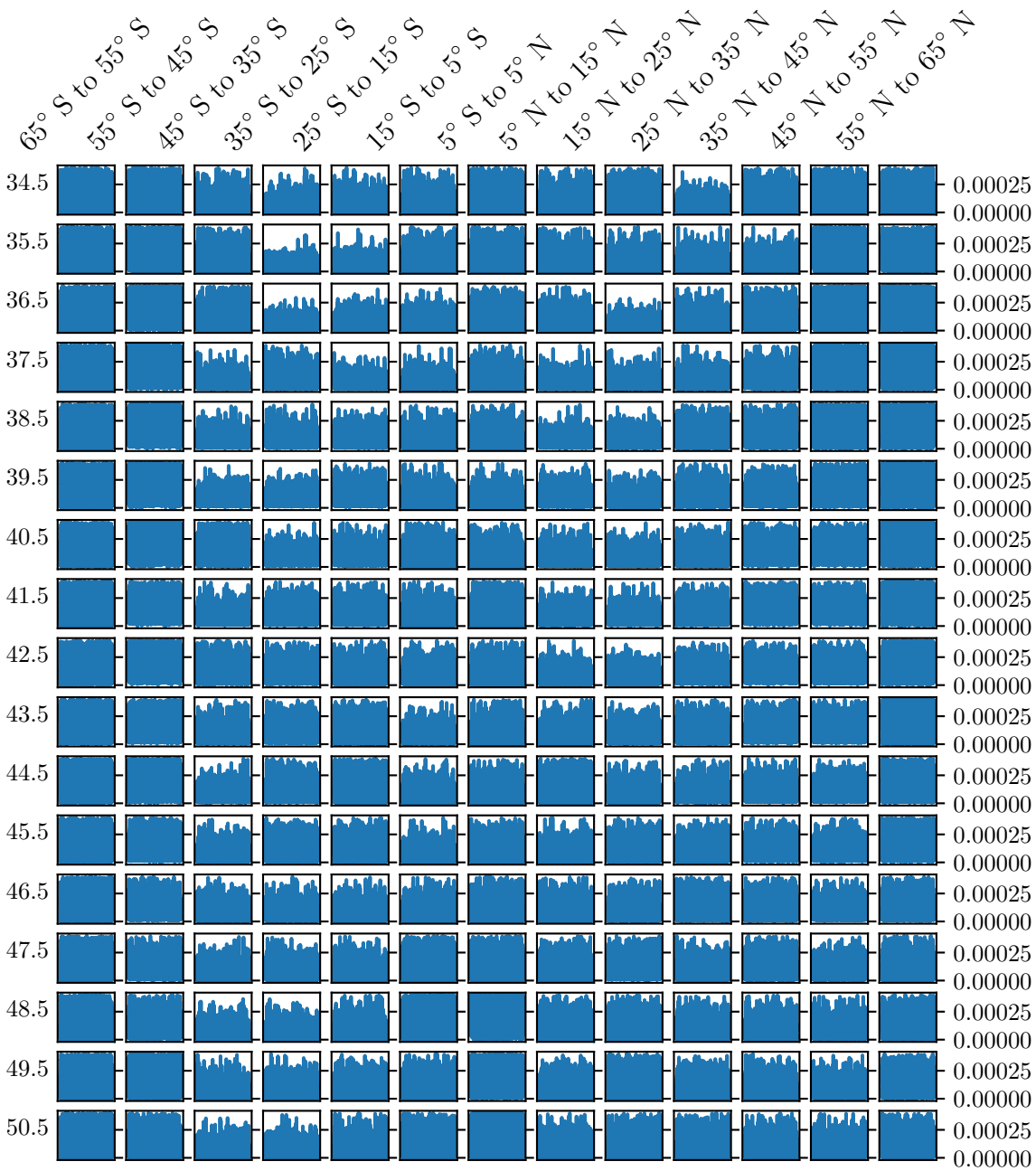


Figure V.4: σ_{trend} Trace Plots Continued.

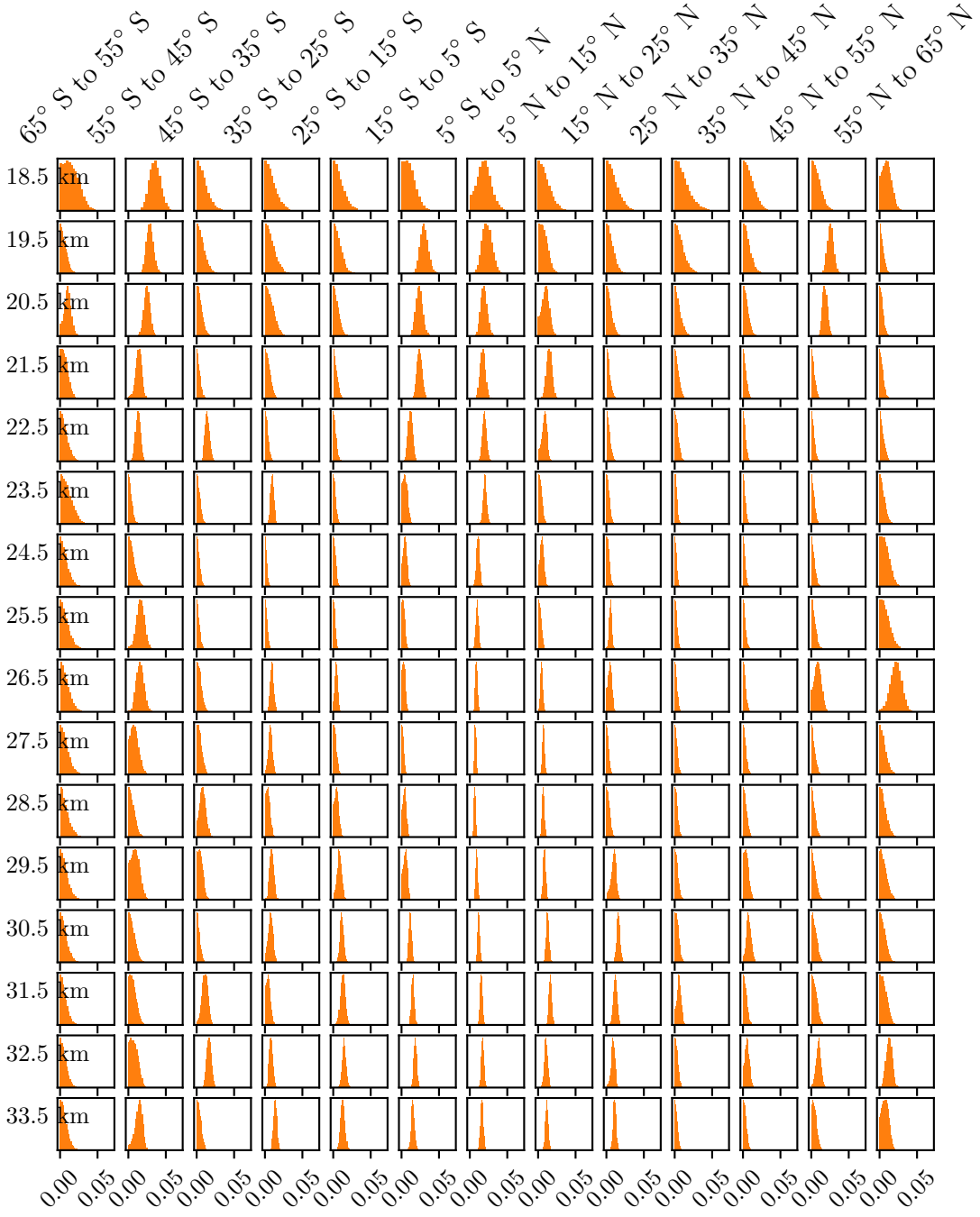


Figure V.5: σ_{AR} Histograms.

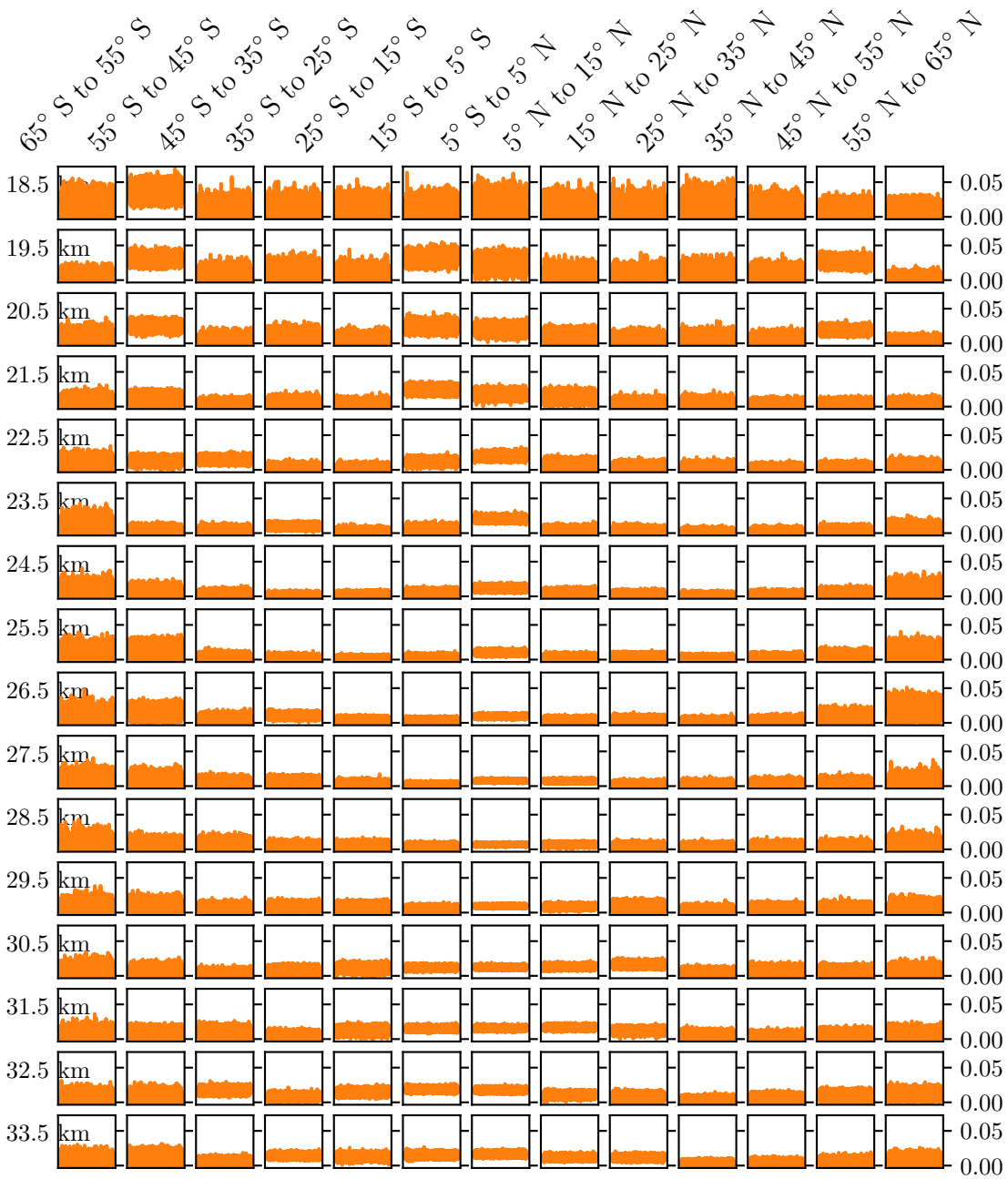


Figure V.6: σ_{AR} Trace Plots.

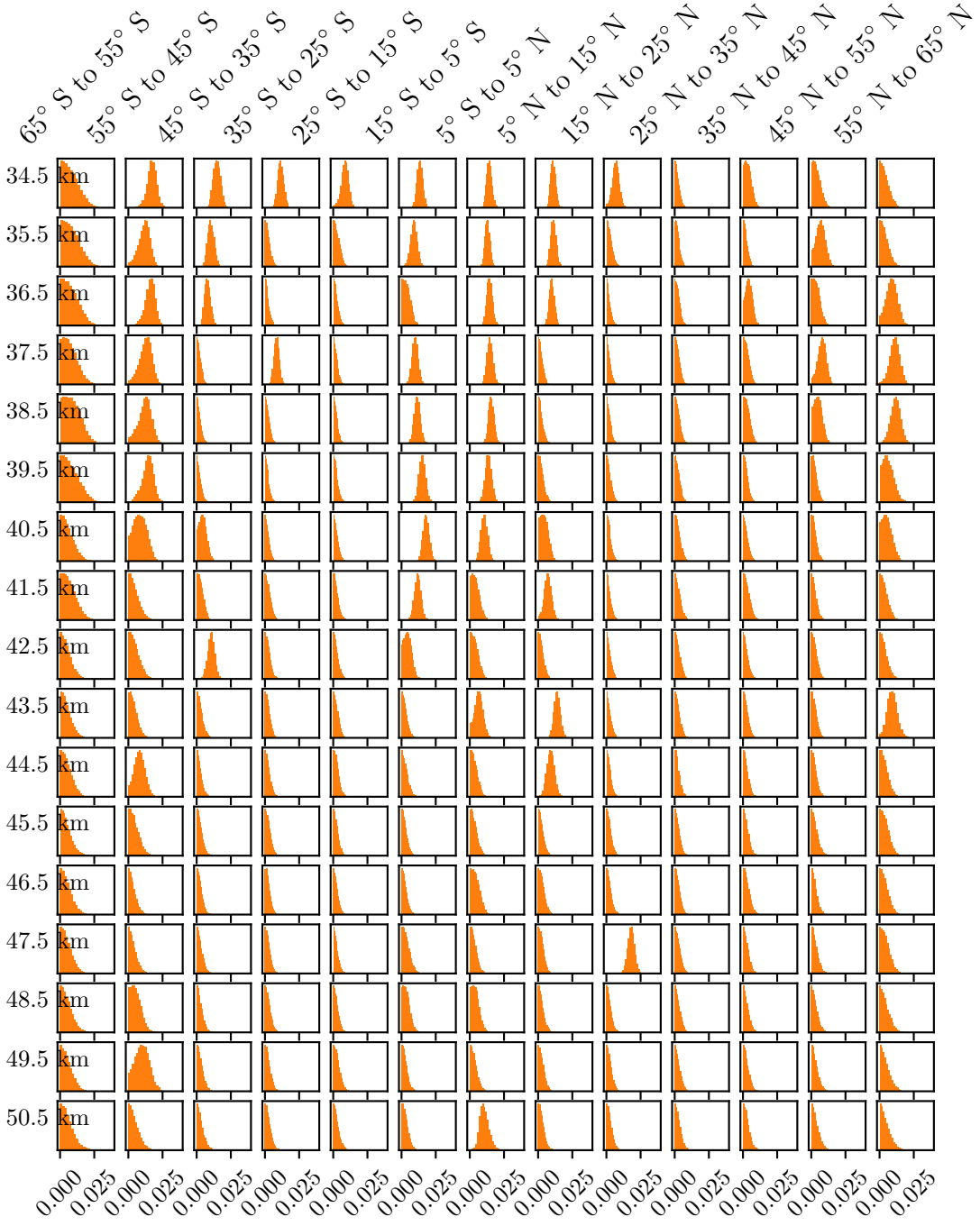


Figure V.7: σ_{AR} Histograms Continued.

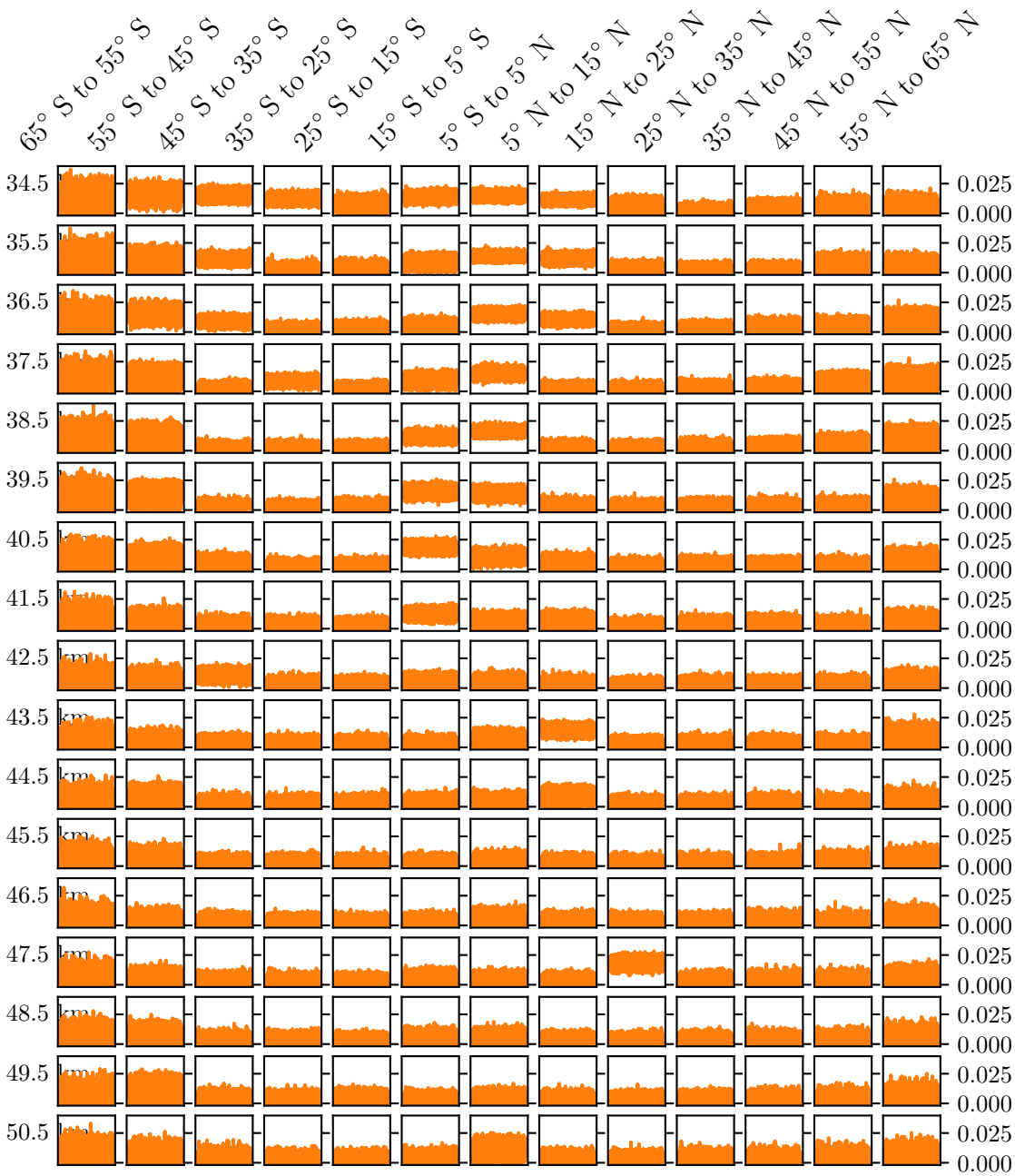


Figure V.8: σ_{AR} Trace Plots Continued.

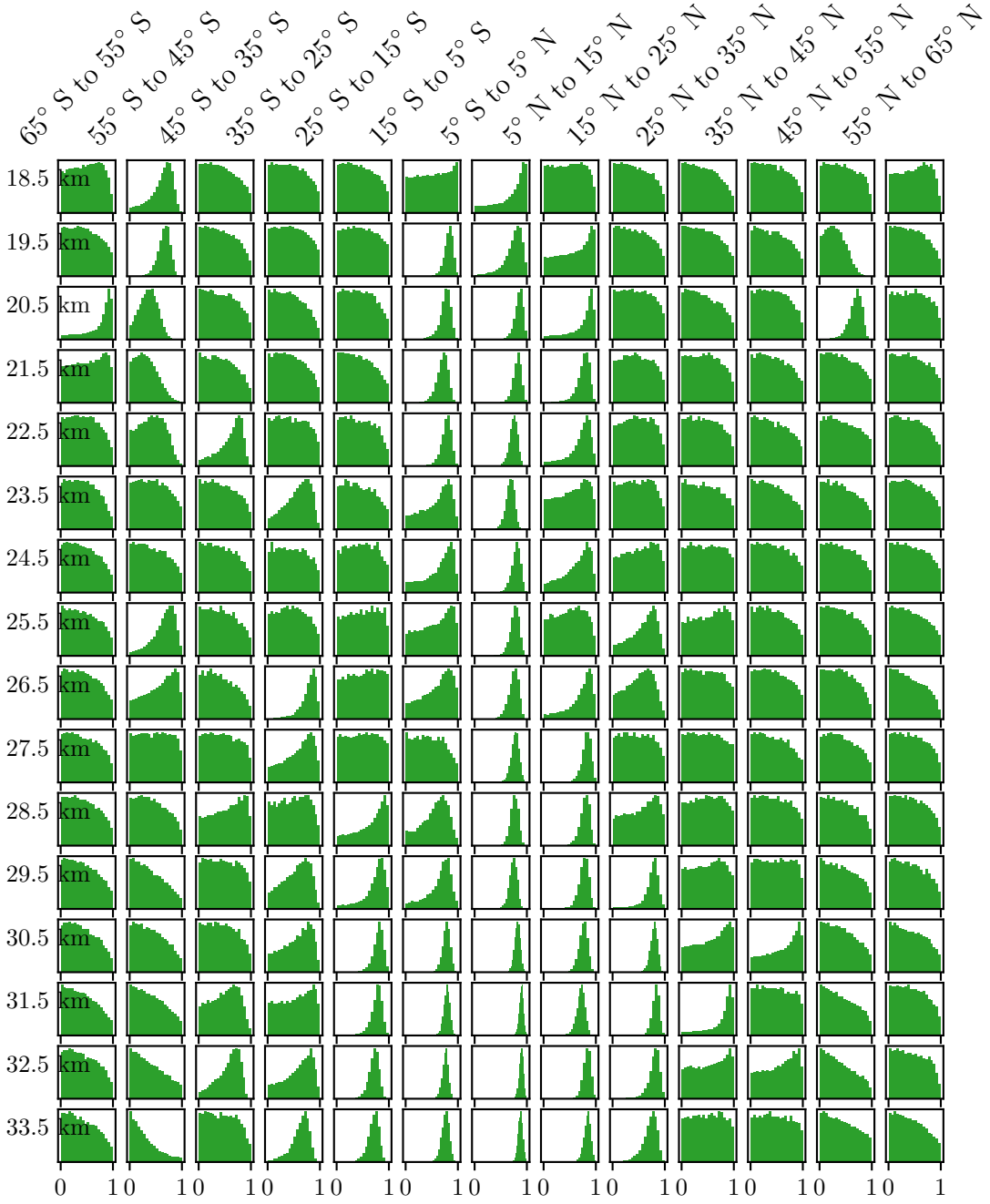


Figure V.9: ρ Histograms.

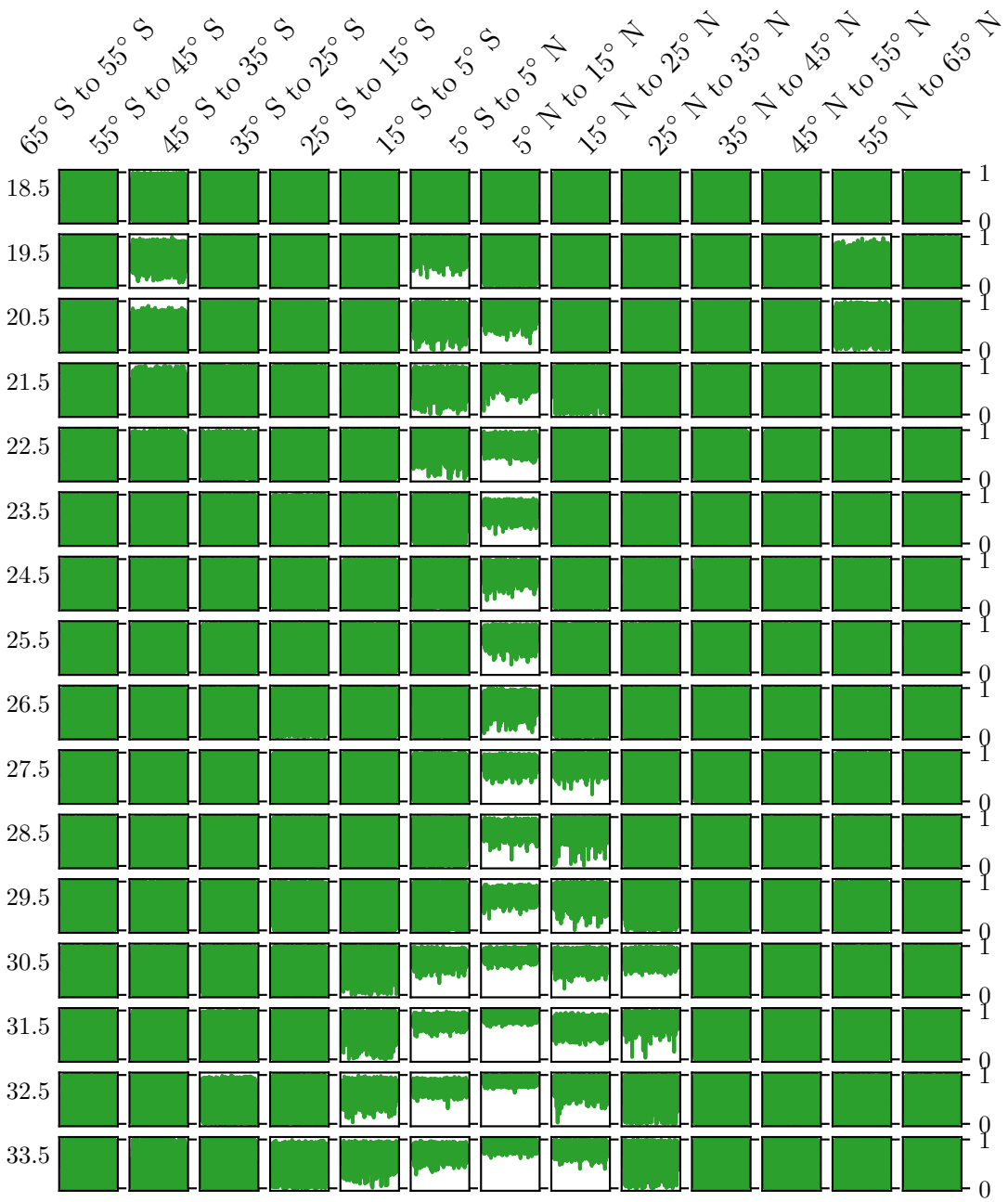


Figure V.10: ρ Trace Plots.

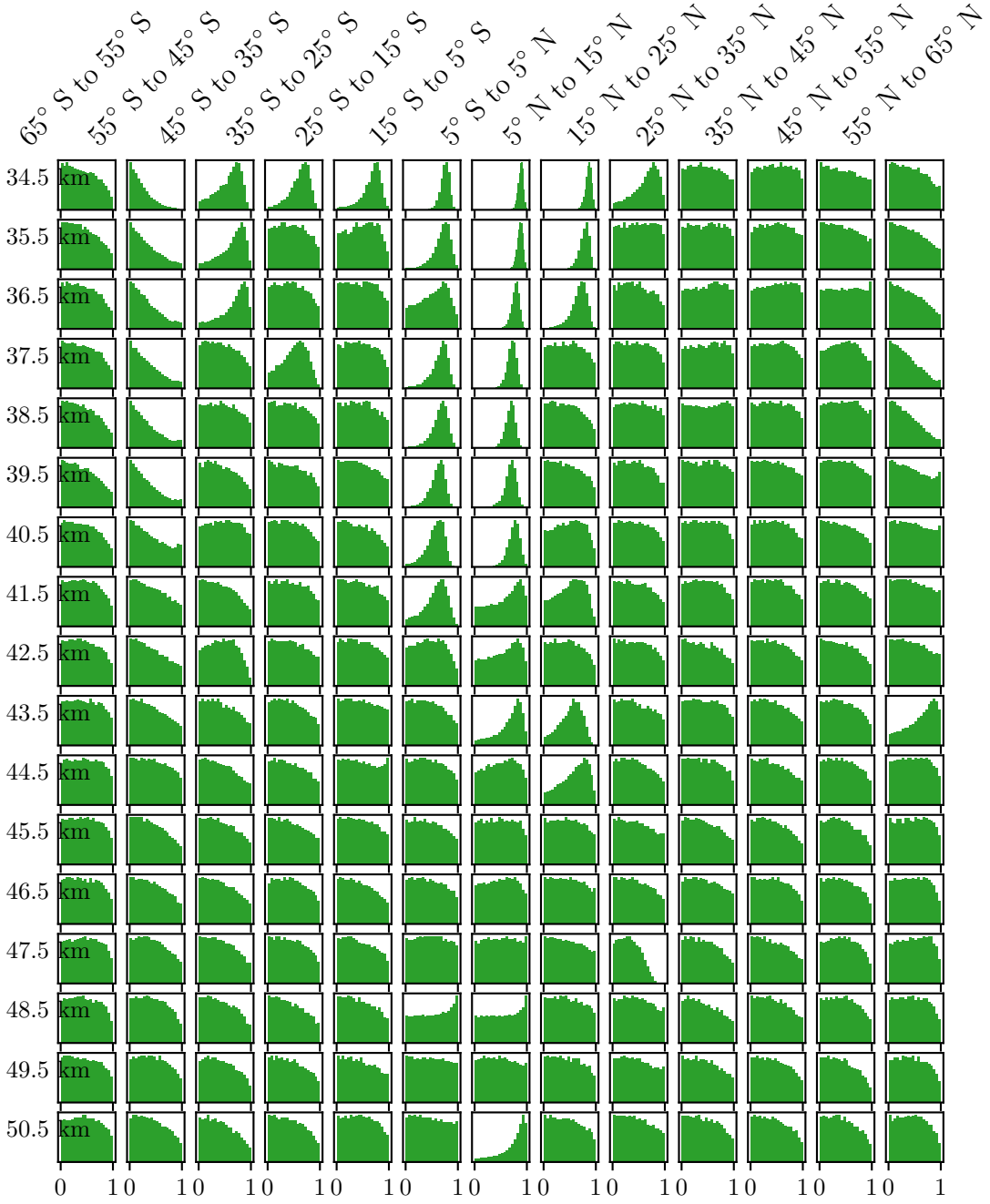


Figure V.11: ρ Histograms Continued.

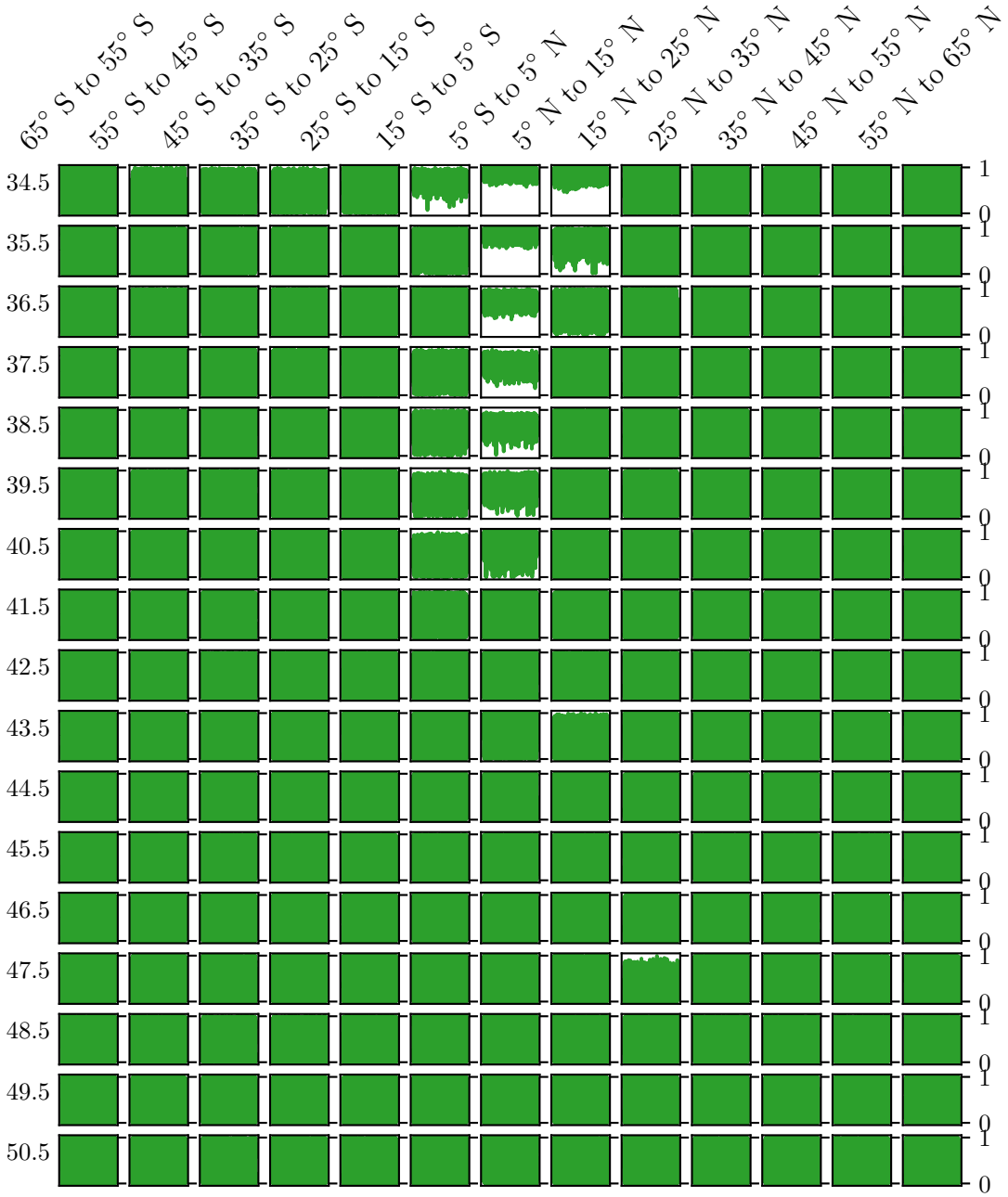
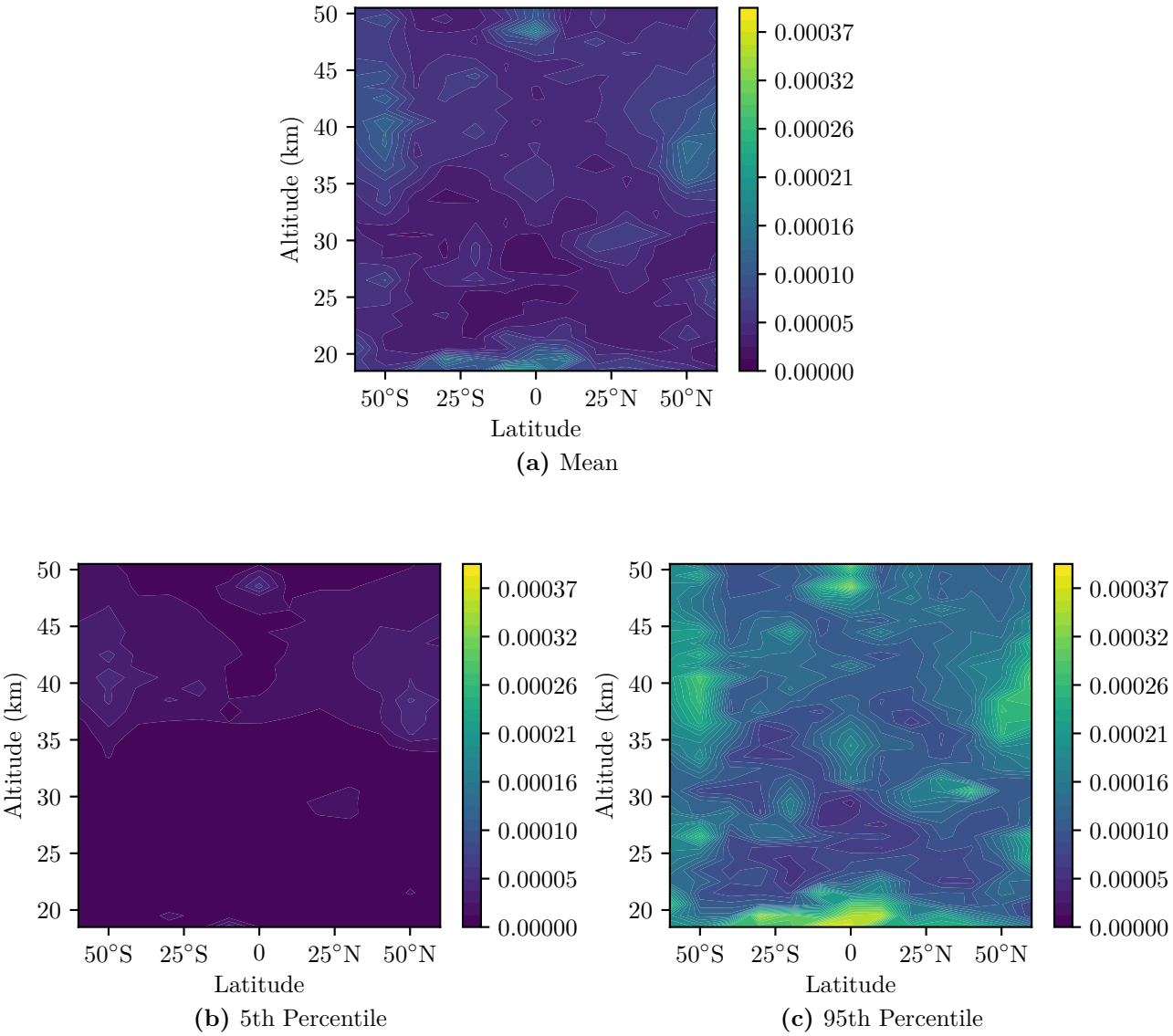


Figure V.12: ρ Trace Plots Continued.

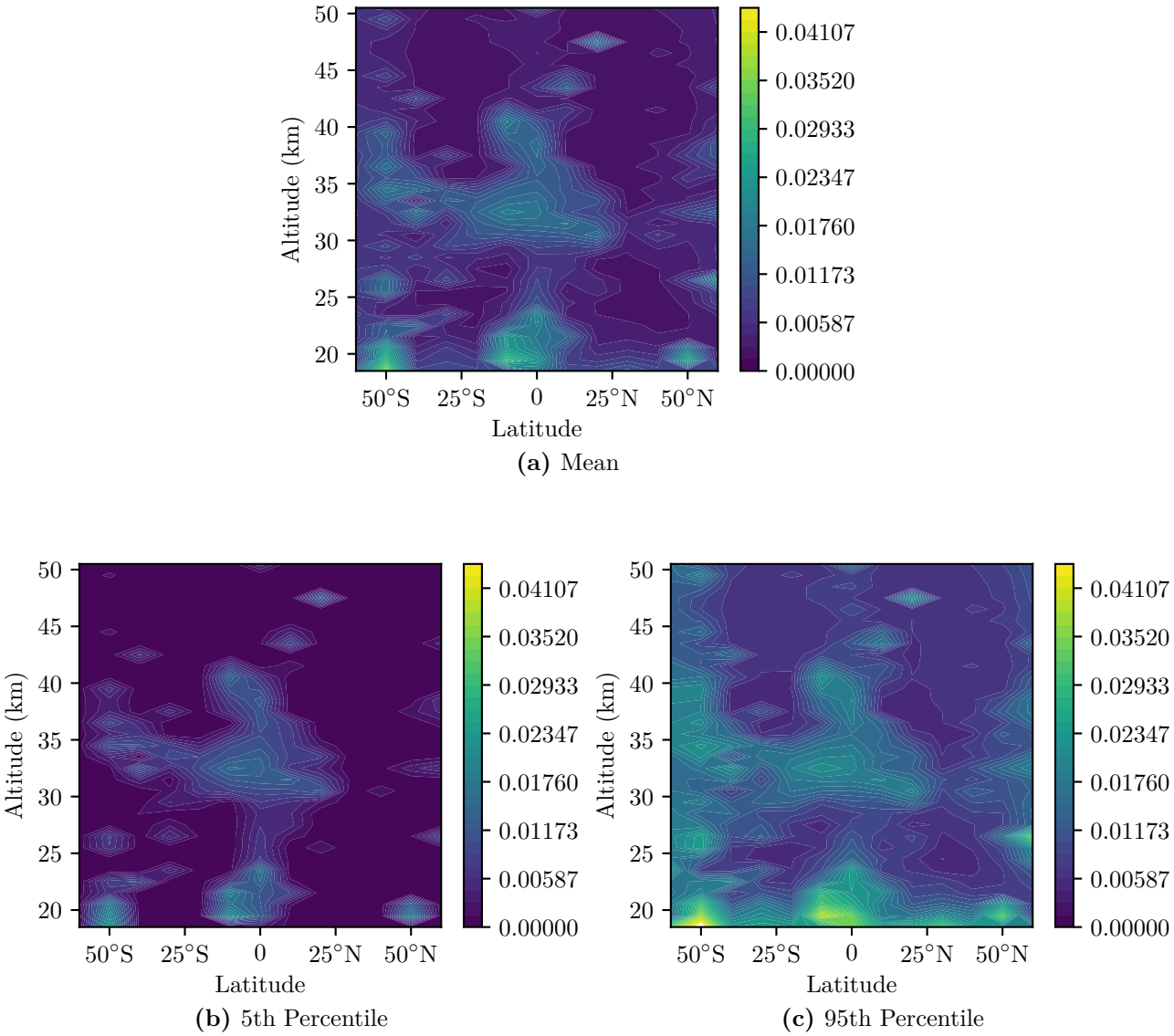


(a) Mean

(b) 5th Percentile

(c) 95th Percentile

Figure V.13: σ_{trend} MCMC Samples.

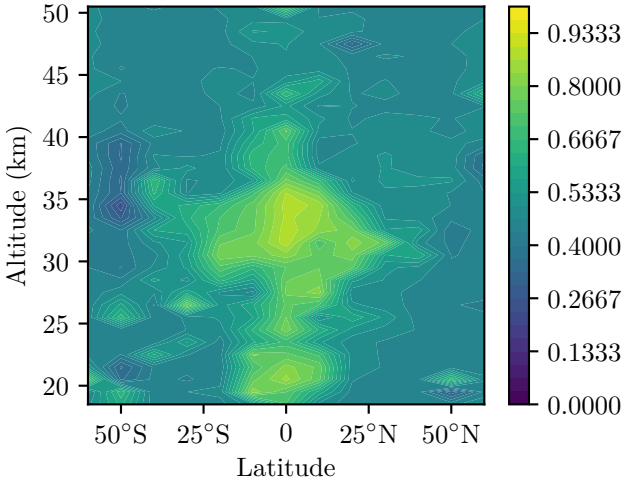


(a) Mean

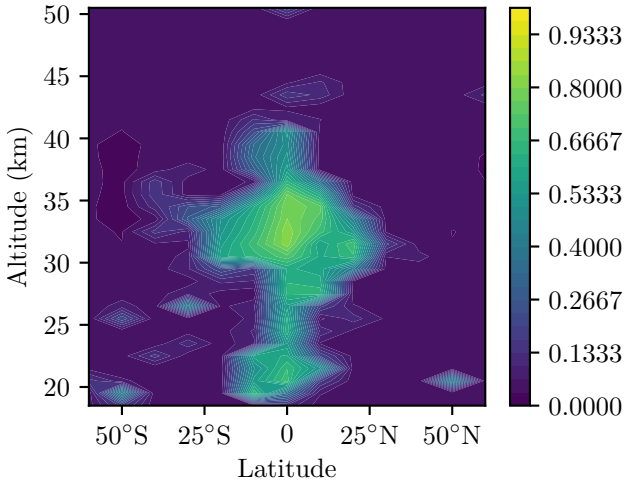
(b) 5th Percentile

(c) 95th Percentile

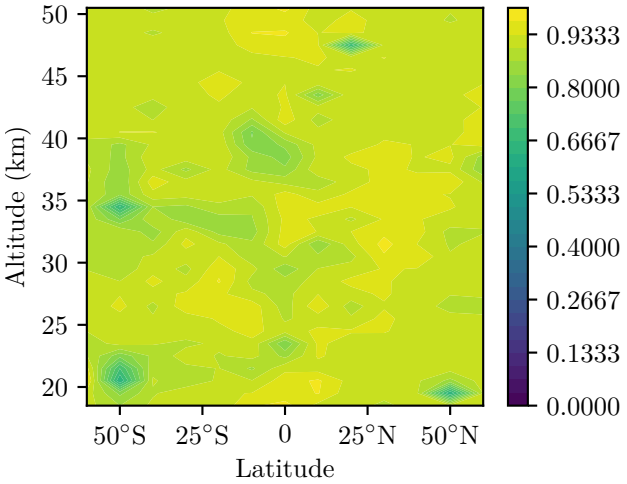
Figure V.14: σ_{AR} MCMC Samples.



(a) Mean

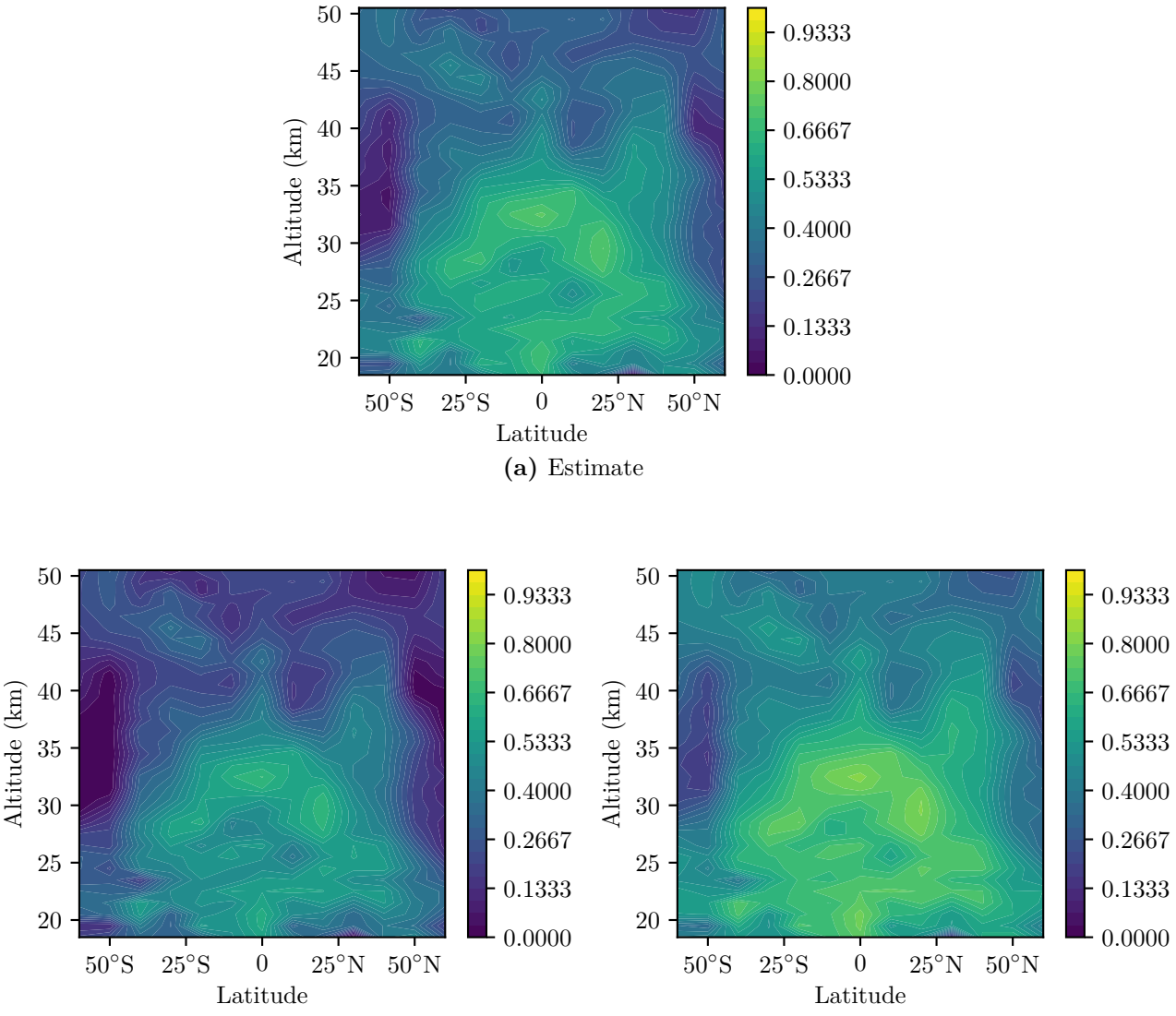


(b) 5th Percentile



(c) 95th Percentile

Figure V.15: ρ MCMC Samples.



(a) Estimate

(b) 95% confidence interval lower bound

(c) 95% confidence interval upper bound

Figure V.16: Estimated ρ from Prais-Winsten Procedure.

Appendix W

Uncorrelated Gaussian Random Vectors are Independent

In this appendix, we prove that if Gaussian distributed random vectors are uncorrelated (zero covariance) then they are also independent. The same also applies to random variables where we consider the random vectors to just be of length 1.

Proof

We define the random vectors $\mathbf{X} = (X_1, ..., X_m)^T$ and $\mathbf{Y} = (Y_1, ..., Y_n)^T$. To prove that these random vectors are independent is to prove that their joint distribution equals the product of their marginal distributions as follows:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}). \tag{W.1}$$

Let us define $\mathbf{Z} = (\mathbf{X}^T, \mathbf{Y}^T)^T$ so that $p(\mathbf{z}) = p(\mathbf{x}, \mathbf{y})$. If $\mathbf{X}$ and $\mathbf{Y}$ are Gaussian distributed, we may show their independence by showing that the following equation:

$$\frac{1}{\sqrt{(2\pi)^{m+n}|\boldsymbol{\Sigma}_z|}} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_z)^T \boldsymbol{\Sigma}_z^{-1}(\mathbf{z}-\boldsymbol{\mu}_z)} = \frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}_x|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_x)^T \boldsymbol{\Sigma}_x^{-1}(\mathbf{x}-\boldsymbol{\mu}_x)} \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}_y|}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}_y^{-1}(\mathbf{y}-\boldsymbol{\mu}_y)} \tag{W.2}$$

holds. Here we have defined $\boldsymbol{\mu}_z$, $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y$, $\boldsymbol{\Sigma}_z$, $\boldsymbol{\Sigma}_x$, and $\boldsymbol{\Sigma}_y$ as the means and covariance matrices of $\mathbf{Z}$, $\mathbf{X}$, and $\mathbf{Y}$ respectively. Now, with $\mathbf{X}$ and $\mathbf{Y}$ as uncorrelated random vectors, we have,

$$\boldsymbol{\Sigma}_z = \begin{bmatrix} \boldsymbol{\Sigma}_x & 0 \\ 0 & \boldsymbol{\Sigma}_y \end{bmatrix}. \tag{W.3}$$

We also have,

$$\boldsymbol{\mu}_z = \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}. \tag{W.4}$$

From noting that the determinate and inverse of $\boldsymbol{\Sigma}_z$ are,

$$|\boldsymbol{\Sigma}_z| = \begin{vmatrix} \boldsymbol{\Sigma}_x & 0 \\ 0 & \boldsymbol{\Sigma}_y \end{vmatrix} = |\boldsymbol{\Sigma}_x||\boldsymbol{\Sigma}_y| \tag{W.5}$$

and

$$\boldsymbol{\Sigma}_z^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_x & 0 \\ 0 & \boldsymbol{\Sigma}_y \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_x^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}_y^{-1} \end{bmatrix} \tag{W.6}$$

from block matrix properties, it can be seen quite readily that Equation W.2 holds. Therefore it is proven that if Gaussian random vectors are uncorrelated then they are also independent.

We also note here that independence of the random vectors $\mathbf{X}$ and $\mathbf{Y}$ implies that

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}), \tag{W.7}$$

and similarly that

$$p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}). \tag{W.8}$$